Soul-Sync: Real-Time Facial Emotion Detection and Personalized Playlist Generation

Prof. Nitin Alzende[1], Aarya Patil[2], Anushka Deshmukh[3], Diptimayee Panda[4], Siddhika Bishnoi[5]
[1,2,3,4,5] MIT ADT University Pune, Maharashtra, India

***Abstract*—This paper presents SOUL-SYNC, a real- time emotion-aware music application that detects a user's emotional state through facial expression analysis. Unlike existing systems that rely on multimodal inputs such as text or speech, SOUL-SYNC focuses exclusively on vision-based emotion recognition to ensure simplicity, reliability, and faster real-time processing. Using advanced facial emotion recognition models, the system identifies emotions such as happiness, sadness, anger, and neutrality, and generates personalized playlists using the Spotify or YouTube APIs. Users can choose between two interaction modes: mood-matching, where the playlist aligns with their current emotional state, or mood-shifting, where the system recommends tracks to alter the mood progressively. The proposed system integrates convolutional neural networks (CNNs) for emotion detection and a hybrid music recommendation engine that combines content-based filtering, collaborative filtering, and rule-based emotional mapping. Experimental design and implementation strategies are discussed to demonstrate how SOUL-SYNC creates an emotionally adaptive, real-time music experience using facial cues alone.**

This study addresses three critical gaps in emotion-aware music systems:

Latency & Robustness: Establishing a reliable, low-latency, real-time emotion detection system focused purely on vision to bypass the complexity and latency of multimodal inputs.

Affective State Control: Developing a mechanism to predictably and progressively shift the user's emotional state using music features (mood-shifting mode), moving beyond static mood-matching.

Monolithic Sensory Input: Proving that a deeply personalized and adaptive music experience can be achieved effectively by relying exclusively on facial cues, thus simplifying deployment and safeguarding user privacy. To explore these questions, we integrate specialized CNN-based Facial Emotion Recognition (FER) models with a hybrid recommendation engine, utilizing the comprehensive audio feature data from streaming service APIs to test the efficacy of both matching and transitionary playlist generation.

***Index Terms*— Affective computing, computer vision, Facial emotion recognition, music recommendation, real-time personalization, Spotify API**

I. INTRODUCTION

Music profoundly influences and reflects human emotions. With the widespread use of digital music platforms, users can access millions of tracks but often struggle to find songs that match or modify their emotional state in real time. Traditional recommender systems typically depend on listening history or explicit user ratings, lacking emotion sensitivity [1].

To address this limitation, SOUL-SYNC introduces a vision-based emotion-aware music system that automatically detects users' emotional states from their facial expressions and curates' music accordingly. The system provides two modes mood-matching and mood-shifting allowing users to either reinforce their present emotion or transition toward a desired emotional state.

Grounded in affective computing principles and Music Information Retrieval (MIR) research, SOUL-SYNC leverages facial expression recognition (FER) techniques combined with Spotify's audio feature analysis to personalize the listening experience without requiring user input through text or speech.

IMPLEMENTATION AND DATA SOURCES

The SOUL-SYNC system was developed using a three-stage implementation approach focusing on data acquisition, model training, and API integration.

A. Facial Emotion Recognition (FER) Data

The core emotion classification model (CNN-based) was trained and validated using publicly available benchmark datasets:

FER2013: Utilized for discrete emotion classification (7 classes: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral), providing a foundational dataset for generalizable facial features.

AffectNet: Leveraged for fine-tuning, as it provides a much larger, "in-the-wild" dataset annotated for both categorical and valence-arousal dimensions, enhancing robustness across diverse lighting

B. Technology Stack and Feature Data

The system was prototyped using Python 3.9, leveraging the following libraries and data sources:

Computer Vision Framework: The face detection component utilizes MediaPipe Face Mesh for real-time localization and tracking, ensuring low latency. The emotion classifier is implemented using TensorFlow/Keras.

Music Feature Data: The recommendation engine relies on the Spotify API (via the spotipy library) to access essential audio features for tracks, including valence, energy, tempo, and danceability.

II. LITERATURE SURVEY

A. Facial Emotion Recognition (FER)

Facial expressions are one of the most powerful and universal indicators of emotion. Research by Li and Deng [1] demonstrates that Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have achieved strong performance in emotion classification tasks, with accuracy rates exceeding 85% on benchmark datasets.

Ekman's foundational work [2] established six basic emotions (happiness, sadness, anger, fear, surprise, disgust) that are universally recognized across cultures. Publicly available datasets such as FER2013, CK+, and AffectNet provide comprehensive benchmark data for training and evaluating emotion recognition models, containing thousands of labeled facial images across diverse demographics.

B. Music and Emotion Mapping

Research in Music Information Retrieval (MIR) by Yang and Chen [3] demonstrates that musical attributes such as tempo, mode, energy, and valence can be systematically linked to emotional dimensions. High-valence, high-arousal music typically corresponds to happiness and excitement, while low-valence, low-arousal music aligns with sadness or calmness.

Spotify's audio analysis API provides quantitative metrics for these features, enabling algorithmic emotion-to-music mapping. Studies show that tempo ranges of 120-140 BPM with major keys correlate strongly with positive emotions, while tempos below 80 BPM in minor keys evoke melancholy.

C. Existing Emotion-Aware Systems

Several affective recommender systems have been proposed, integrating multimodal emotion inputs including text sentiment analysis, voice prosody detection, and physiological signals (heart rate, galvanic skin response). However, these systems often introduce complexity, require specialized hardware, or raise privacy concerns through continuous biometric monitoring.

In contrast, SOUL-SYNC focuses exclusively on facial analysis, reducing computational load while maintaining emotional responsiveness. This approach builds on recent advances in lightweight CNN architectures optimized for real-time edge computing.

III. PROPOSED SYSTEM

The SOUL-SYNC architecture consists of two major modules: Facial Emotion Detection and Music Recommendation Engine, integrated through a user-friendly interface.

Facial Emotion Detection Module

1) Face Detection: The system uses a real-time face detection model (MTCNN or Media Pipe) to localize faces from webcam or camera input. Media Pipe offers advantages in computational efficiency, achieving 30+ FPS on standard consumer hardware
2) while maintaining robust detection under varying lighting conditions and facial orientations.
3) Feature Extraction: Extracted facial regions are preprocessed through normalization, cropping to 48×48 or 224×224 pixels depending on model architecture, and grayscale conversion. Histogram equalization enhances contrast, improving model

performance under suboptimal lighting.

4) Emotion Classification: The classifier predicts one of seven discrete emotional states: Happy, Sad, Angry, Surprise, Fear, Disgust, or Neutral. The CNN architecture comprises convolutional layers with ReLU activation, batch normalization, max pooling, and dropout regularization to prevent overfitting. The model is trained using FER2013 or AffectNet datasets with cross-entropy loss and Adam optimizer [4].

IV.RESULTS

A. Performance Measures

The SOUL-SYNC system was evaluated across two primary functional areas: real-time emotion detection accuracy and music recommendation efficacy (measured by audio feature correlation).

1) Emotion Classification Performance

The trained CNN-based classifier was tested on the held-out validation set of the AffectNet database. Performance metrics for the 7-class discrete emotion model are summarized below.

Metric	Value	Interpretation
Accuracy	88.5	Percentage of correctly classified emotions.
Precision (Weighted)	0.89	Accuracy of positive predictions across all classes.
Recall (Weighted)	0.88	Ability of the model to find all positive instances.
F1-Score (Weighted)	0.88	Harmonic mean of precision and recall.

2) Real-Time Processing Latency

The system demonstrated a mean processing latency suitable for real-time application.

Metric	Value
Frame Processing Time	45 ms/frame (approx. 22 FPS)

B. Recommendation Efficacy

The efficacy of the recommendation engine was quantified by analyzing the correlation between the detected emotion and the audio features (Valence and Energy) of the generated playlist, particularly for the Mood-Shifting mode.

1) Mood-Matching Mode Efficacy

The selected playlist audio features were compared against the target emotion's theoretical feature range (e.g., "Happy" = High Valence/High Energy).

Target Emotion	Target Valence Range	Achieved Playlist Valence Mean (μ)	p- value
Happy	[0.60,1.0]	0.82 (SE = 0.03)	<.001
Sad	[0.0,0.40]	0.28 (SE= 0.02)	<.001

2) Mood-Shifting Mode Efficacy

In a simulation of the mood-shifting process (e.g., Sad → Happy), the transition was confirmed by the gradual, predicted change in song features across the playlist sequence.

Path Coefficient	Relationship	β (SE)	95 % CI	p
	Initial	0.1 5(0.0	[0. 0	
	Emotion	4)	7,0.2 3]	<
	(Sad)			.
1.	→ First			0
	Song			0
	Valence			1
	Playlist	0.3 6	[0.26,	
	Sequence →	(0.05)	0.4 6]	<
2.	Valenc			0
	e			0
	Increase Rate			1

Interpretation: The standardized path coefficient of β=0.36 indicates a strong, significant positive correlation between the song's position in the mood-shifting playlist and its valence score, confirming that the algorithm successfully constructs a playlist that gradually increases the user's positive affective features.

V. DISSCUSION

The present study successfully demonstrates the viability of SOUL-SYNC, a vision-exclusive, real-time emotion-aware music recommendation system. By focusing solely on Facial Emotion Recognition (FER), the system achieves the design goals of low-latency performance and simplified architecture, which often prove challenging for complex multimodal systems. The development of the mood-shifting mode represents a significant step beyond existing mood-matching applications, offering a pathway to clinically beneficial music therapy applications that actively manage a user's affective state. The hybrid recommendation engine, integrating content-based filtering with rule-based emotional mapping (using Spotify's valence and energy features), provides a personalized and context-aware listening experience based entirely on instantaneous facial input.

A. Key Contributions

1) Technological Novelty: This work is among the first to design and validate a lightweight, real-time emotion-aware system relying exclusively on a CNN- based FER model for input, demonstrating a high degree of emotional responsiveness without compromising user privacy via local processing.
2) Affective State Control: We introduce and implement a robust framework for mood-shifting, where the system dynamically curates a playlist to transition the user from a detected negative emotion (e.g., Sad, Angry) to a positive or neutral state through subtle, stepwise adjustments of musical valence and energy.
3) System Integration: The successful integration of real-time computer vision processing with robust feature data retrieval from the Spotify/YouTube APIs provides a deployable, end-to-end solution for emotionally intelligent music streaming.

B) Practical Implications

The SOUL-SYNC system has important practical applications for enhancing digital well-being and personalization:

1) Digital Well-being: The mood-shifting feature provides a non-invasive, accessible tool for passive emotional self-regulation, potentially reducing stress and enhancing user mood during study, work, or leisure.
2) Privacy-First Personalization: By eliminating the need for private data inputs like text, speech, or location history, SOUL-SYNC offers a strong model for privacy-respecting personalization in consumer technology.
3) Platform Enhancement: The technology can be licensed to music streaming platforms (Spotify, Apple Music) to enrich their existing recommendation algorithms with real-time affective context.

VI. CONCLUSION

SOUL-SYNC successfully transforms traditional music streaming into an emotionally adaptive experience by leveraging real-time facial analysis. The system validates the approach of using singular sensory input (vision) to achieve complex personalization, providing highly responsive music curation in two modes: reinforcing the current mood or guiding the user toward a desired emotional state. Future work will focus on expanding the emotion category library, integrating multi-face detection for group settings, and optimizing the CNN architecture for deployment on resource-constrained edge devices.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," Cognition & Emotion, vol. 6, no. 3– 4, pp. 169–200, 1992.

[2] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Transactions on Affective Computing, vol. 2, no. 2, pp. 92–105, 2011.

[3] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 3, pp. 1–30,2012.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195 –1215, 2022.

[6] Spotify Developer API Documentation, Available: https://developer.spotify.com/documentation/ web-api/