Research Trends International Journal of Technology and Innovation
An International Peer-Reviewed, Refereed & Open-Access Journal
ISSN : [ISSN]
Available Online at :
https://rtijti.com/
doi : https://doi.org/[DOI-PREFIX]/RTIJTI

Emotion-Driven Music Recommendation System Using Multimodal Data

Chetan Devle , Prof. Shelke S. A

Department of Entc Engineering, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted : 06 August 2026
Published : 08 August 2026

Publication Issue
Volume 01, Issue 03
July-September 2026

Page Number
11-17

ABSTRACT

Recommending music that matches or regulates a listener's emotional state requires knowing that state first, and most commercial recommenders skip this step entirely, relying instead on listening history and collaborative filtering that say nothing about how the listener feels right now. This paper presents a multimodal emotion-driven recommendation system that infers a listener's affective state from three signals captured during a session: facial expression from the device camera, sentiment extracted from any text the listener enters (search queries, chat, or captions), and, where available, vocal tone from a short voice sample. Each modality is scored independently by a lightweight classifier and the three scores are combined through a confidence-weighted late fusion step, so a modality that is unavailable or low-confidence in a given session, such as a blank camera feed, does not silently distort the result. The fused output is a position on a two-dimensional valence-arousal plane rather than a single discrete label, which is then matched against a track catalogue annotated with the same valence-arousal coordinates derived from audio features such as tempo, energy, and mode. We describe the fusion architecture, the mapping from emotional state to track features, and a feedback loop that adjusts future recommendations when a listener skips a suggested track quickly. We evaluate the system on a constructed listening-session dataset for classification accuracy per modality, fusion accuracy against self-reported mood, and recommendation acceptance rate compared to a collaborative-filtering baseline.

Index Terms - Emotion Recognition, Multimodal Fusion, Music Recommendation, Valence-Arousal Model, Affective Computing, Session-Based Feedback

I. INTRODUCTION

1.1 Background and Motivation

Music streaming platforms already know a great deal about what a listener has played before, but very little about how that listener feels in the moment they open

the app. Collaborative filtering and content-based recommenders both optimize for what is statistically likely to be played next given past behaviour, which is a different question from what would suit a listener who is anxious before an exam, low-energy after a long shift, or unexpectedly happy. Affective computing has made it feasible to estimate a person's emotional state from readily available signals, facial expression, text, and voice, but most music recommenders still treat this signal as out of scope.

Combining these two ideas, an emotion estimate feeding a recommendation step, is not new in isolation, but doing it robustly requires handling the fact that no single modality is reliably available in every session. A listener may have their camera off, may not type anything, and may not want to record their voice, so a practical system has to degrade gracefully when one or more inputs are missing rather than requiring all three.

1.2 Problem Statement

We set out to build a recommendation system with three properties. First, it must estimate emotional state from whichever modalities are actually available in a given session, without requiring all of facial video, text, and voice to be present. Second, it must express that estimate in a form that maps naturally onto measurable audio properties of a track, rather than a small fixed set of discrete labels that force a coarse match. Third, it must adapt when a listener's behaviour, such as skipping a recommended track within a few seconds, indicates the estimate or the mapping was wrong for that session.

1.3 Contributions

This paper makes four contributions. We present a confidence-weighted late-fusion architecture that combines facial-expression, text-sentiment, and vocal-tone classifiers into a single estimate while tolerating missing or low-confidence modalities. We adopt a continuous valence-arousal representation for the fused emotional estimate instead of a discrete emotion label, and describe how track-level audio features are mapped into the same two-dimensional space so that matching becomes a nearest-neighbour problem. We define a lightweight session feedback loop that treats a fast skip as a negative signal on the current valence-arousal target and adjusts subsequent recommendations within the same session. We evaluate the system on a constructed listening-session dataset for per-modality classification accuracy, fused-estimate accuracy against self-reported mood, and recommendation acceptance rate relative to a collaborative-filtering baseline.

1.4 Organisation of the Paper

Section II reviews emotion recognition modalities, valence-arousal models, and existing mood-based recommenders. Section III sets out the system architecture. Section IV details the methodology: per-modality classification, fusion, the valence-arousal mapping, and the feedback loop. Section V covers the implementation. Section VI presents the evaluation. Section VII discusses the design trade-offs, Section VIII states the limitations, and Section IX concludes with future work.

II. RELATED WORK

2.1 Facial Expression and Physiological Emotion Recognition

Facial-expression recognition using convolutional networks is well established for discrete categories such as happy, sad, angry, and neutral, typically trained on labelled datasets of posed or spontaneous expressions [1]. Physiological signals such as heart-rate variability and galvanic skin response have also been used to estimate arousal, but require wearable hardware that most listeners will not have attached during casual listening [2], so we treat facial expression as the primary visual modality and do not require physiological sensing.

2.2 Text and Speech Sentiment

Sentiment analysis on short text has moved from lexicon-based scoring to transformer-based classifiers that handle negation and context reasonably well [3]. Vocal-tone emotion recognition, sometimes called speech emotion recognition, extracts prosodic features such as pitch variance and speaking rate rather than word content, and is generally reported as less accurate

than facial or text modalities in isolation, which is consistent with our decision to weight it by confidence rather than treat it as equally reliable [4].

2.3 The Valence-Arousal Model

The circumplex model of affect represents emotion as a point in a continuous two-dimensional space defined by valence, how positive or negative a state is, and arousal, how calm or energetic it is [5]. This representation is attractive for a recommendation system because audio features such as tempo, loudness, and mode have been empirically linked to listener-perceived valence and arousal, giving a natural shared coordinate system between an emotional estimate and a track's properties [6].

2.4 Mood-Based Music Recommendation

Existing mood-based recommenders generally take one of two approaches: asking the listener to self-report a mood from a fixed list, or inferring mood purely from listening history and skip behaviour without any direct affective signal [7]. Systems that do use a camera or microphone typically rely on a single modality and a discrete label, which forces a coarse match to a handful of pre-tagged playlists rather than a continuous, per-session adjustment.

2.5 Gap Addressed

The gap we address is the combination of multimodal, confidence-weighted fusion with a continuous valence-arousal target and a session-level feedback loop. Prior mood-based recommenders that use a live signal generally commit to one modality and one discrete label per session; we instead treat each modality as an independent, gracefully-degrading estimator and use the fused continuous output to drive nearest-neighbour matching against track features, adjusting within the session as skip behaviour provides additional evidence.

III. SYSTEM ARCHITECTURE

3.1 Overall Structure

The system has two paths that run continuously during a listening session: an emotion-estimation path, which collects whatever modalities are available and produces a valence-arousal estimate, and a recommendation path, which matches that estimate against a track catalogue and adjusts based on skip feedback. Both paths run per session rather than once at login, so the estimate can shift as the session continues.

3.2 Component Overview

Three per-modality classifiers, facial, text, and vocal, each output an independent valence-arousal estimate with an associated confidence score, and produce nothing for a modality that has no input in the current session rather than a default guess. The fusion module combines whichever estimates are present, weighted by their confidence, into a single fused valence-arousal point. The recommendation engine holds a track catalogue pre-annotated with valence-arousal coordinates and returns the nearest unplayed tracks to the fused point. The feedback module watches playback events and, on a fast skip, nudges the target point away from the skipped track's coordinates for the remainder of the session.

3.3 Data Flow

At the start of a session, the client begins capturing whichever inputs the listener has permitted. Each active modality periodically emits a valence-arousal estimate with a confidence score to the fusion module, which recomputes the fused point on every update and passes it to the recommendation engine. The recommendation engine returns a ranked list of nearest tracks, and playback events are streamed back to the feedback module, which can revise the fused target before the next recommendation is requested.

IV. METHODOLOGY

4.1 Facial Expression Classification

Facial frames captured from the device camera are passed through a convolutional classifier trained to output a valence-arousal point rather than a discrete category, using a regression head on top of a standard face-emotion backbone. A face-detection confidence check runs first; if no face is detected for a sustained period, the modality reports no estimate rather than a

stale or default value, which is what allows the fusion step to safely exclude it.

4.2 Text Sentiment Classification

Any text the listener enters during the session, search queries, typed captions, or chat, is scored by a transformer-based sentiment classifier fine-tuned to output valence directly and arousal indirectly from lexical intensity cues such as capitalisation, punctuation, and intensifiers. Because text is sparse in a typical session, this modality's confidence is scaled down for short inputs and decays over time since the last text event, so an old message does not dominate a later fused estimate.

4.3 Vocal Tone Classification

Where a listener opts in to a short voice sample, prosodic features, pitch, energy, and speaking rate, are extracted and passed to a lightweight classifier trained on speech-emotion datasets to produce a valence-arousal estimate. This modality is given the lowest default confidence weight of the three, consistent with the comparatively lower reliability reported for speech emotion recognition in isolation, and is only used when the listener explicitly enables microphone access.

4.4 Confidence-Weighted Fusion

The fusion module computes the fused valence-arousal point as a weighted average of the available modality estimates, where each modality's weight is the product of a fixed reliability prior for that modality and its own reported confidence for the current input. A modality with no current estimate contributes a weight of zero rather than a default value, so a session with only a camera feed is driven entirely by the facial estimate, and a session with all three inputs blends them proportionally to their confidence.

4.5 Valence-Arousal to Track Mapping

Each track in the catalogue is annotated with a valence-arousal coordinate derived from audio features, mode and tempo contributing to valence and arousal respectively, energy and loudness contributing mainly to arousal, following the mapping conventions established in prior work on audio-feature-based mood estimation. Recommendation is then a nearest-neighbour search in this two-dimensional space around the fused listener estimate, with a diversity constraint that avoids returning near-duplicate tracks in the same result set.

4.6 Session Feedback Loop

A skip within a short threshold of playback start is treated as a negative signal on the recommended track's coordinates: the feedback module nudges the session's effective target point away from the skipped track's position by a fixed step, before the next recommendation request. A track played to completion or explicitly liked nudges the target slightly toward that track's coordinates instead, so the session-level target drifts based on accumulated behaviour rather than only the raw fused estimate.

4.7 Evaluation Method

We evaluated three things on a constructed listening-session dataset: per-modality classification accuracy against labelled ground truth for each of the three classifiers in isolation; fused-estimate accuracy against self-reported mood collected at intervals during simulated sessions; and recommendation acceptance rate, measured as the proportion of recommended tracks played past the skip threshold, compared against a collaborative-filtering baseline with no emotional input.

V. IMPLEMENTATION

5.1 Technology Stack

The facial and vocal classifiers are implemented in PyTorch and served through a lightweight inference API; the text sentiment classifier uses a fine-tuned transformer served through the same API layer. The client is a web application built in React that requests camera, microphone, and text input permissions independently, so a listener can grant any subset. The recommendation engine and track catalogue are served from a Node.js backend backed by a relational database storing track metadata and precomputed valence-arousal coordinates.

5.2 Real-Time Inference Pipeline

Camera frames are sampled at a low rate, sufficient for expression tracking without materially increasing bandwidth or battery use, and sent to the inference API for scoring; text events are scored as they are entered rather than batched, so a typed query can influence the very next recommendation. Each inference call returns both an estimate and a confidence score, which the client forwards to the fusion module running in the backend alongside the current session state.

5.3 Track Catalogue and Feature Precomputation

Valence-arousal coordinates for each track are precomputed offline from standard audio features and stored alongside the track record, so recommendation-time lookups are a nearest-neighbour search over precomputed points rather than a live audio-analysis step. The catalogue is indexed to support efficient nearest-neighbour queries so that recommendation latency does not grow materially as the catalogue size increases.

5.4 Privacy and Data Handling

Camera frames and voice samples are processed for inference and discarded rather than stored, and only the resulting valence-arousal estimate and confidence score are retained for the duration of the session. Permissions for camera, microphone, and text-based inference are requested and can be revoked independently, and a listener who grants none of them still receives recommendations, falling back to the collaborative-filtering baseline with no emotional weighting.

VI. RESULTS AND EVALUATION

6.1 Per-Modality Classification Accuracy

On the constructed evaluation set, the facial-expression classifier was the most accurate single modality against labelled ground truth, followed by the text-sentiment classifier, with the vocal-tone classifier trailing both, consistent with the relative reliability priors we assigned during fusion design. All three modalities performed noticeably worse on borderline, low-intensity expressions than on clearly expressed emotional states, which is expected given that ambiguous inputs are inherently harder to classify regardless of modality.

6.2 Fused-Estimate Accuracy Against Self-Reported Mood

The confidence-weighted fused estimate tracked self-reported valence and arousal more closely than any single modality taken alone, including in sessions where one modality was missing entirely, which supports the design choice of excluding absent modalities rather than substituting a default value. Sessions using all three modalities showed the closest agreement with self-report, while single-modality sessions showed a wider spread, most noticeably when the only available modality was vocal tone.

6.3 Recommendation Acceptance Rate

Recommendations driven by the fused valence-arousal estimate were played past the skip threshold more often than recommendations from the collaborative-filtering baseline with no emotional input, and the gap was largest in sessions where the listener's mood, by self-report, differed noticeably from their typical listening history, which is exactly the case a purely history-based recommender is least equipped to handle.

6.4 Effect of the Session Feedback Loop

Enabling the skip-based feedback loop improved acceptance rate further within a session compared to using the static fused estimate for every recommendation in that session, indicating that behavioural correction within a session captures information the upfront multimodal estimate alone does not, such as a listener's reaction to how a track actually sounds once played.

VII. DISCUSSION

The central design choice in this system is to let modalities be absent rather than forcing every session into a fixed input requirement. A recommender that insists on camera, text, and voice before it will personalise anything will simply not be used by most listeners in most sessions; treating each modality as

optional and weighting by confidence is what makes multimodal emotion sensing practical for a mainstream listening product rather than a lab demonstration.

A second point worth making explicitly is that a continuous valence-arousal target is a better fit for music matching than a small set of discrete emotion labels, because two listeners both labelled as happy can have very different arousal levels, one wanting upbeat energetic tracks and the other wanting calm, pleasant ones, and a discrete label collapses that distinction while a two-dimensional target preserves it.

VIII. LIMITATIONS

This work has four limitations we want to state plainly. The evaluation was run on a constructed listening-session dataset rather than on a live deployment with real listeners over extended periods, so acceptance-rate figures may not generalise directly to production usage patterns. Facial-expression and vocal-tone classifiers were trained on general-purpose emotion datasets rather than data collected specifically during music listening, and expressions during passive listening may differ from the more expressive data such classifiers are typically trained on. The valence-arousal mapping for tracks relies on audio features as a proxy for perceived mood, which is a well-established but imperfect approximation, and does not account for lyrical content, which can meaningfully affect perceived valence independent of tempo and mode. Finally, the system has not been evaluated for demographic or cultural bias in facial-expression or vocal-tone classification, which is a known concern in affective computing generally and warrants dedicated evaluation before any production deployment.

IX. CONCLUSION AND FUTURE WORK

9.1 Conclusion

We presented a multimodal emotion-driven music recommendation system that estimates a listener's affective state as a continuous valence-arousal point from whichever of facial expression, text sentiment, and vocal tone are available in a session, fuses them by confidence rather than requiring all three, and matches the fused estimate against audio-feature-derived track coordinates, adjusting within the session as skip behaviour provides feedback. On a constructed evaluation dataset, the fused estimate tracked self-reported mood more closely than any single modality, and recommendations driven by it were accepted more often than a collaborative-filtering baseline, particularly when a listener's momentary mood diverged from their typical history.

9.2 Future Work

Future work includes a live pilot with real listeners over an extended period to measure acceptance rate and retention under actual usage rather than constructed sessions, collection of music-listening-specific facial and vocal emotion data to retrain the per-modality classifiers on more representative expressions, incorporation of lyrical sentiment into the track-side valence-arousal mapping, and a dedicated evaluation of demographic and cultural bias across the facial and vocal classifiers before any production deployment.

REFERENCES

[1] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," Neural Networks, vol. 64, pp. 59-63, 2015.

[2] R. W. Picard, "Affective Computing," MIT Press, Cambridge, MA, 1997.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.

[5] J. A. Russell, "A Circumplex Model of Affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.

[6] Y. E. Kim et al., "Music Emotion Recognition: A State of the Art Review," in Proc. 11th Int. Society for Music Information Retrieval Conf. (ISMIR), 2010.

[7] X. Wang, D. Rosenblum, and Y. Wang, "Context-Aware Mobile Music Recommendation for Daily Activities," in Proc. 20th ACM Int. Conf. Multimedia, 2012.