

CrossMood

Music Emotion Recognition Across Cultures

When you search a music website for songs to help you relax, the system needs to identify music that listeners find calm. Yet the same song may communicate different emotions to people from different cultural backgrounds. Taking these differences into account could help mood search and playlist recommendations better reflect the intended audience.

CrossMood explores how song audio and listener culture can be used to predict a group's emotion ratings. Under the conditions tested in this study, cultural information helped improve prediction, while a lightweight model achieved performance close to that of a more complex model. The project offers an approach to music emotion prediction that accounts for cultural differences.

Research Background

Music is often described as a form of expression that crosses language barriers. Even without understanding the lyrics, listeners may hear excitement, tension, or sadness in a song. Shared responses, however, do not mean that everyone will describe a song with the same words or give it the same ratings. Musical traditions, familiar films and media, and personal experience can all shape interpretation.

Many commonly used music emotion datasets rely mainly on English labels, and some focus on Western music. These resources have advanced the field, but they raise a question: how well do judgments learned from particular audiences and vocabularies apply to other groups? Cross-cultural research, including GlobalMood, has found that even emotion words with direct translations do not always produce the same rating patterns across groups. [1]

This project asks whether a prediction model can account for such differences. For example, given a song and a target listener culture, it estimates how calm, sad, or energetic that group is likely to judge the music. It predicts a group's evaluation of the emotion expressed by a song. It does not read an individual user's current mood or force different cultures to receive different predictions.

Two concepts need to be distinguished. Listener culture refers to the group evaluating a song; song origin refers to where the song comes from. If Korean listeners evaluate a French song, these are Korea and France, respectively. The model uses listener culture while the experiments also examine whether it relies too heavily on musical cues associated with song origin.

Data and Research Approach

The study uses the public GlobalMood dataset, which contains 1,180 songs from 59 countries and ratings from listeners in Egypt, France, South Korea, Mexico, and the United States. Each group evaluates songs using emotion words in its own language, allowing the same music to be described from different cultural perspectives. [1] After matching and screening the available audio, this project used 1,145 songs.

Listener ratings provide the basis for training and evaluation. If a group gives a song a high rating for calmness, the model needs to learn a relationship between the audio and that judgment. These targets come from listener surveys and mainly represent group averages. They do not imply that every person in a group has the same response.

To process sound, a computer first needs to convert audio into numerical features. This study uses the pretrained audio model CLAP for that step. CLAP has learned relationships between sounds and text descriptions, providing audio features for the prediction task. [2] Its parameters remain frozen, meaning they are not updated during this project’s training. The study trains the emotion predictor that follows CLAP and focuses on how it uses cultural information.

The model produces emotion tag scores and a more general pair of emotion coordinates. Tag scores describe details such as how sad or calm a song sounds. The two coordinates, called valence and arousal, or V-A, describe how positive or negative an emotion is and how calm or activated it is. Sadness and peacefulness can both have low arousal while differing in valence.

The V-A training targets in this study are calculated from listener tag ratings and reference positions assigned to the emotion words. They are not separate V-A ratings collected by GlobalMood. Performance on these targets therefore measures how well the model predicts this derived representation; it is not a direct measurement of an individual’s emotional state.

The study compares four approaches: a basic model that uses cultural information, a more compact lightweight model, a complex interaction model that also uses word meanings, and a comparison model that does not use cultural information. This tests both prediction quality and whether additional complexity brings worthwhile gains. It also evaluates training that aims to reduce reliance on song-origin cues when handling music from unfamiliar regions.

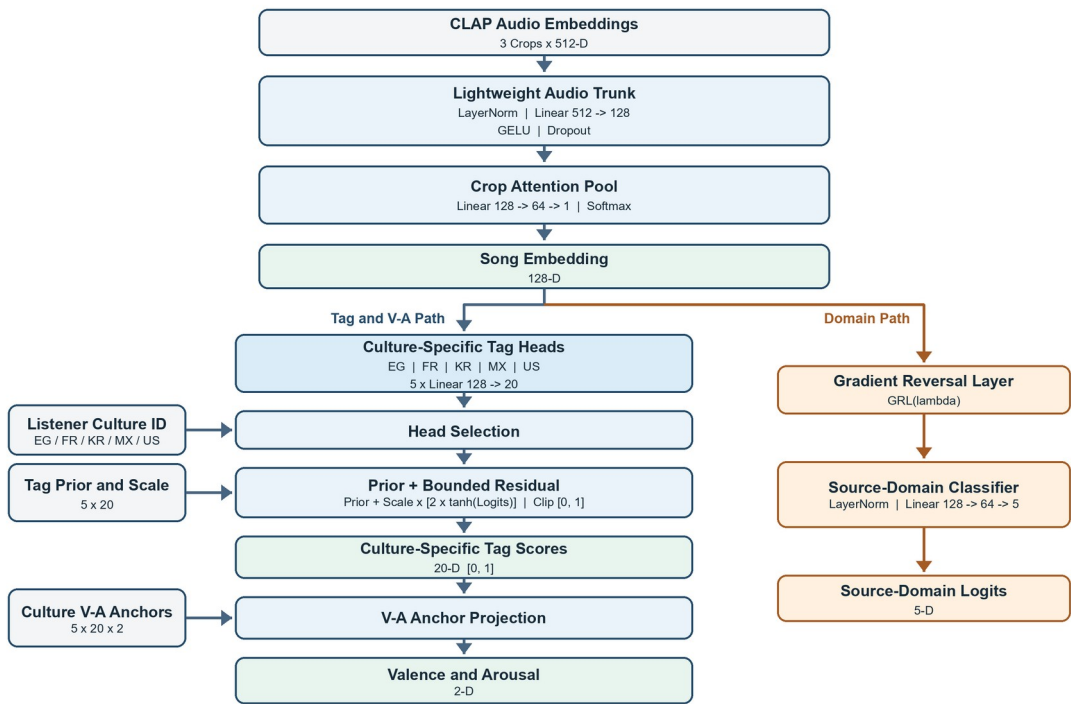


Figure 1. Simplified overview of CrossMood Lite. Frozen CLAP extracts audio features, and the predictor combines audio information with listener culture to estimate emotion tag scores. Culture-specific mapping converts these scores into valence and arousal. The dashed domain adversarial branch is optional and is used only during training to reduce reliance on song-origin cues.

Experiments and Main Findings

A useful evaluation asks whether the model can handle songs it has not learned during training. The study splits data by song, keeping all cultural ratings of a song in the same partition. This prevents the model from encountering a test song through another group's ratings. A separate evaluation holds out an entire song-origin region to test performance on music from an unseen region.

Evaluation considers how far predictions are from listener ratings, whether the model captures differences in emotion strength across songs, and whether prediction errors vary greatly between listener groups. These measures answer different questions. Results from the two evaluation settings are interpreted separately rather than ranked together.

First, in the fixed song-split experiment, the trained culture-aware models performed better overall than the reference methods selected for the study. Those references included predicting cultural average ratings, directly matching pretrained audio and text features, and using traditional acoustic features. This suggests that training for the cross-cultural rating task helps the models use information from individual songs.

Second, in the cultural-information comparison, the model using listener culture predicted rating targets more closely than the model without that information. The error gap between listener groups also became smaller. These results support using culture under the tested data and evaluation conditions. They do not show that culture explains every difference in musical emotion, or that a country or language label fully describes a person's cultural identity.

Another main finding concerns model size. CrossMood Lite's prediction network uses fewer trainable parameters while achieving overall performance close to that of the complex interaction model. At the current data scale, additional word-meaning interactions did not consistently improve results. This supports starting with a compact predictor and testing whether further complexity provides a practical benefit.

Here, lightweight refers mainly to the prediction network after CLAP. New audio still requires feature extraction. The speed and memory use of the complete process need separate measurements; a smaller prediction network alone does not establish fast performance on every device.

Training designed to reduce reliance on song origin provided only limited additional gains. Some measures improved slightly, but prediction errors did not improve consistently. The current results therefore do not establish a broad improvement in prediction across origin regions. This finding helps identify where further development still needs evidence.

Potential Applications

CrossMood's direct output is a music emotion prediction model that could support mood search and recommendation features. Suppose a listener wants calm music. A website could combine the target listener culture with predicted emotion scores to filter or rank candidate songs. The judgment of whether a song fits "calm" could then draw on ratings relevant to that audience.

A similar approach could help creators select music for films or short videos. A creator who wants a soundtrack to convey warmth might consult predictions for different cultural audiences, then make a decision using listening tests and audience feedback. Such applications require their own evaluation, and model predictions should support human judgment.

The project also provides a way to include different cultural interpretations of music in a technical system. A service intended for diverse audiences needs to recognize both shared responses and differences. Learning from multiple listener groups can reduce reliance on one group's judgments as a substitute for everyone else's, allowing more cultural perspectives to be represented.

Individual choice remains important. Cultural information can guide an initial prediction, while personal feedback should allow the service to adjust it. A future application could let users choose whether to use a cultural perspective and refine the results through listening and preferences. Group averages should not become fixed assumptions about individuals.

The experiments evaluate emotion rating prediction, not whether people prefer playlists generated from those predictions. A complete recommendation service must also consider personal taste and listening context. The next step toward an application is to test whether better rating predictions lead to song selections that better meet listeners' needs.

Limitations and Future Work

The study currently covers five listener cultures and cannot represent the full range of cultural backgrounds. People within the same region also differ in age, musical experience, and personal preferences. The available group labels do not capture all these factors and should be treated as limited descriptions rather than complete accounts of identity.

Data quality also affects the conclusions. Song-origin regions have unequal sample sizes, and some audio excerpts may contain silence or speech. Weaker performance for a region may reflect both data quantity and audio quality. It should not automatically be interpreted as an inability to understand that region's music.

Future work will prioritize audio screening and add ratings from more listener cultures, alongside a more balanced selection of song origins. With expanded data, the study can reassess when lightweight and complex models are most useful. Measuring the full audio-processing pipeline will also help establish suitable devices and deployment conditions.

Small listening studies could test specific tasks such as finding relaxing music, comparing song selections with and without cultural information. The key question would be whether the playlists better match the target listeners' judgments. Participants should also be able to express opinions that differ from the group average. This would connect model evaluation with actual use.

CrossMood demonstrates a workable approach to predicting music emotion across the listener cultures studied, with a compact model retaining much of the predictive performance. It provides a technical basis for considering listener culture in music services. Broader benefits still need to be tested using data from more cultures and feedback from real use.

References

- [1] Lee H, Çelen E, Harrison P, et al. GlobalMood: A cross-cultural benchmark for music emotion recognition. arXiv:2505.09539, 2025.
- [2] Elizalde B, Deshmukh S, Al Ismail M, et al. CLAP: Learning audio concepts from natural language supervision. ICASSP, 2023.