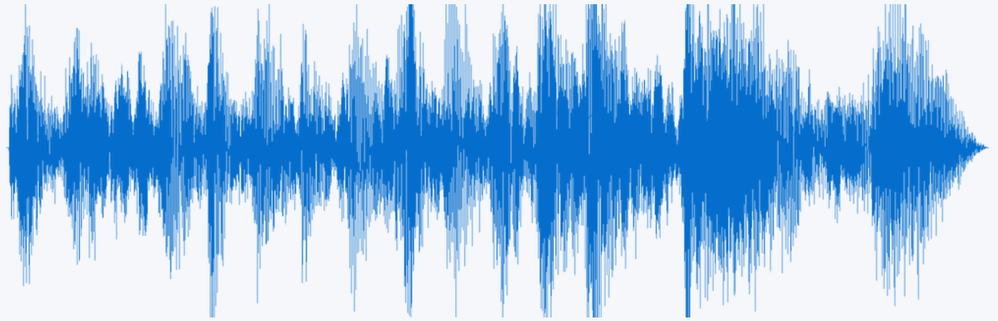
New York Institute of Technology

Emotion-Based Audio Prediction Using Beat-Level Timbre Data

A Machine Learning Approach

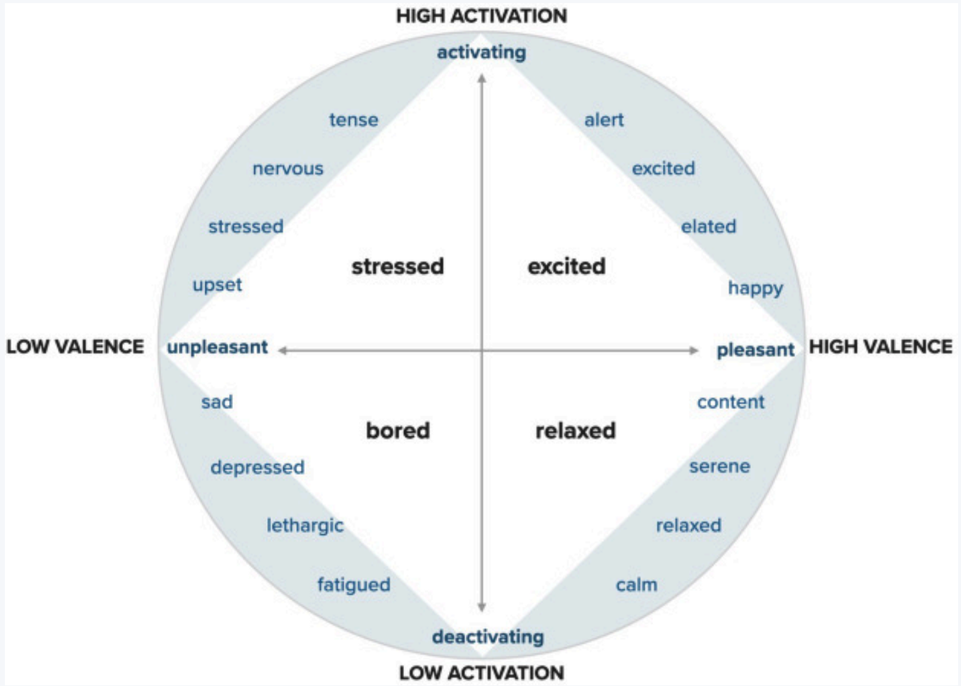


Angel Terrero

Table of Contents

Introduction

- Music can evoke distinct emotional responses, shaping how we interpret and connect with sound.
- Traditional emotion analysis in music relies heavily on lyrics, metadata, or user tags.
- This project explores the potential of beat-level timbre data to classify emotional states without any lyrical input.
- Focus: Classify songs into emotional quadrants based on valence and arousal (Russell's Circumplex Model).



Background and Problem Statement

- Music emotion recognition (MER) traditionally depends on external labels, listener input, or lyrics.
- EchoNest (now part of Spotify) enables detailed per-beat audio analysis using features like timbre, pitch, and rhythm.
- Timbre is a perceptual quality of sound that helps distinguish instruments and texture — it's critical in emotional interpretation.
- The Free Music Archive (FMA) is a large, open-source dataset designed for machine learning in music.
- Can we classify music into emotional quadrants using only beat-level timbre, without any lyrics, tags, or genre labels?
- If so, how accurate can we get using traditional models like Random Forest?

Data Overview

Raw Audio Files

- Start with music tracks in standard audio formats (MP3, WAV, FLAC).

EchoNest Audio Analysis

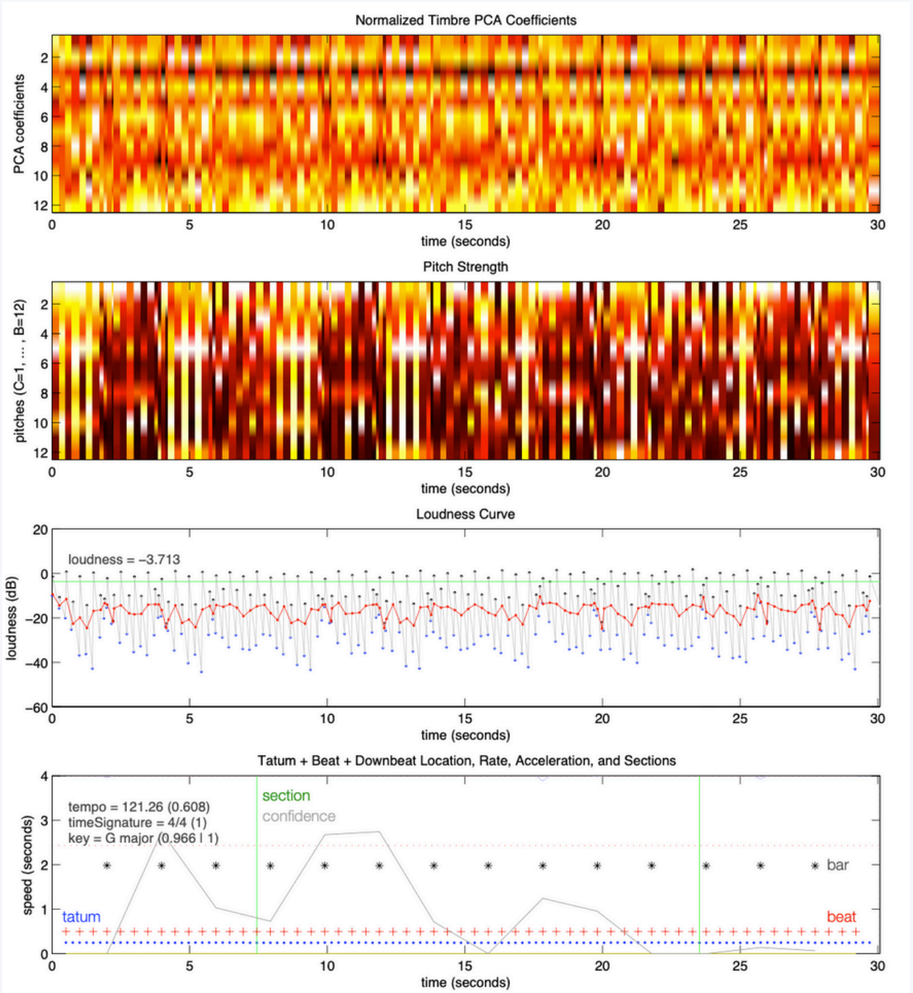
- Uses proprietary "machine listening" techniques based on psychoacoustics and adaptive learning.
- Analyzes music as humans perceive it—at microsecond precisionAnalyzeDocumentation.

Extracted Audio Features

- Segments: Breaks down audio into short segments, each described by:
- Timbre (sound texture or color)
- Pitch (melodic content)
- Loudness (intensity/amplitude)
- Beat-level Analysis: Captures repeated rhythmic patterns (beats), assigning each beat a numeric timbre vector (via PCA)AnalyzeDocumentation.

Numeric Representation

- Timbre represented by a PCA vector of 12 coefficients per beatAnalyzeDocumentation.
- PCA coefficients represent psychoacoustic properties (brightness, flatness, attack).



Plot of the JSON data for a 30-second excerpt of “around the world” by Daft Punk.

Data Cleaning and Preprocessing

Initial Dataset Structure

- Raw EchoNest data (~13,000 tracks × ~250 columns)
- Multi-indexed columns, significant missingness (many null values)

Cleaning Steps

- Flattened multi-level columns for readability
- Removed columns exceeding 23% missing values threshold (3,000 nulls)
- Dropped rows containing any remaining null values

Final Clean Datasets

- echonest_audio_features.csv: Global track-level features (valence, energy, tempo, etc.) – 8 columns, 13,129 rows
- echonest_audio_temporal.csv: Beat-level timbral PCA data – 224 columns (beats), 13,129 rows
- Data structured clearly by track_id for easy alignment

Inputs

raw file	rows × cols	note
`echonest.csv`	~13 k × ~250	multi-index header, many nulls

Outputs

clean file	rows × cols	note
`echonest_audio_features.csv`	13 129 × **8**	neat global features
`echonest_audio_temporal.csv`	13 129 × **224**	$beat_0$ … $beat_{223}$ timbre PCA

Exploratory Data Analysis (EDA)

Quadrant Creation (Valence & Energy)

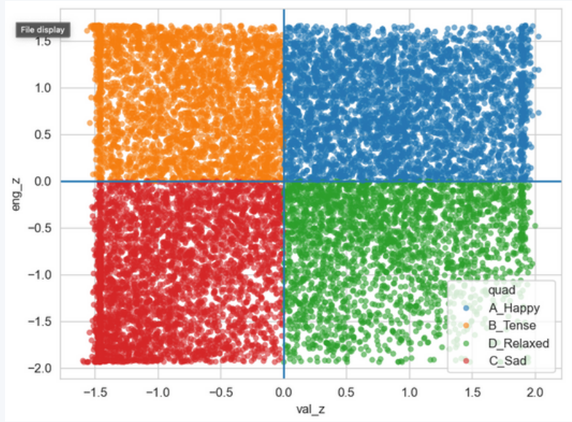
- Used Russell's Circumplex Model to form emotional quadrants based on valence and energy.
- Histogram Insights:
- Valence skewed slightly low; energy roughly uniformo1_Data_Exploration.
- Scatterplot Visualization:
- Clearly defined four emotional quadrants (Happy, Tense, Sad, Relaxed) from standardized valence and energy valueso1_Data_Exploration.

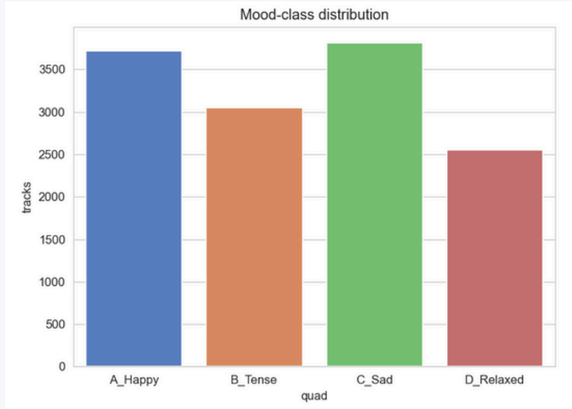
Class Distribution

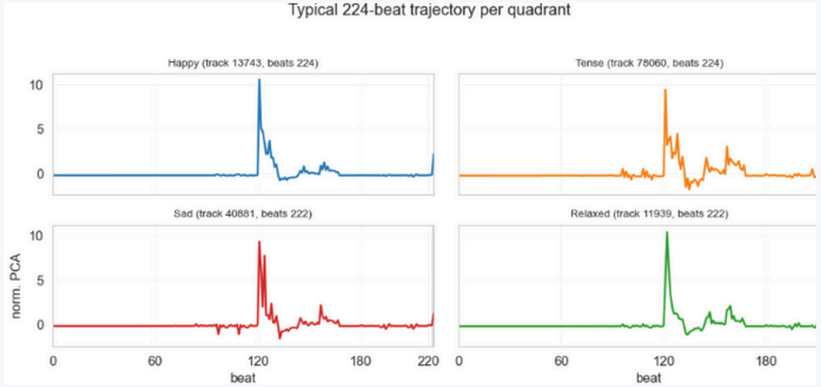
- Slight imbalance observed:
- Sad (29%), Happy (28%), Tense (23%), Relaxed (19%)o1_Data_Exploration.
- Bar Chart clearly illustrating class proportions and imbalance.

Temporal Patterns (Beat-level Analysis)

- Beat-level PCA trajectories differ across emotional quadrants:
- Typical spike at beat ~120 (likely indicating chorus).
- Distinct decay shapes observed for each emotional quadrant, highlighting potential predictive powero1_Data_Exploration.
- Beat-length distribution:
- Most songs end around beats 210–224, ensuring minimal paddingo1_Data_Exploration.







Modeling Approach

Random Forest Overview

- Ensemble model that builds multiple decision trees on random subsets of data and features.
- Combines predictions from many simple models (trees) for robust final decisions.

Input Data

- Features: 224 beat-level PCA timbre coefficients per track.
- Labels: Four emotional categories (Happy, Tense, Sad, Relaxed).

Training & Validation Strategy

- Stratified 5-fold Cross-validation:
- Dataset split into 5 folds with balanced emotional classes.
- Iteratively trains on 4 folds, validates on the 5th; repeats five times to ensure stability.

Handling Class Imbalance

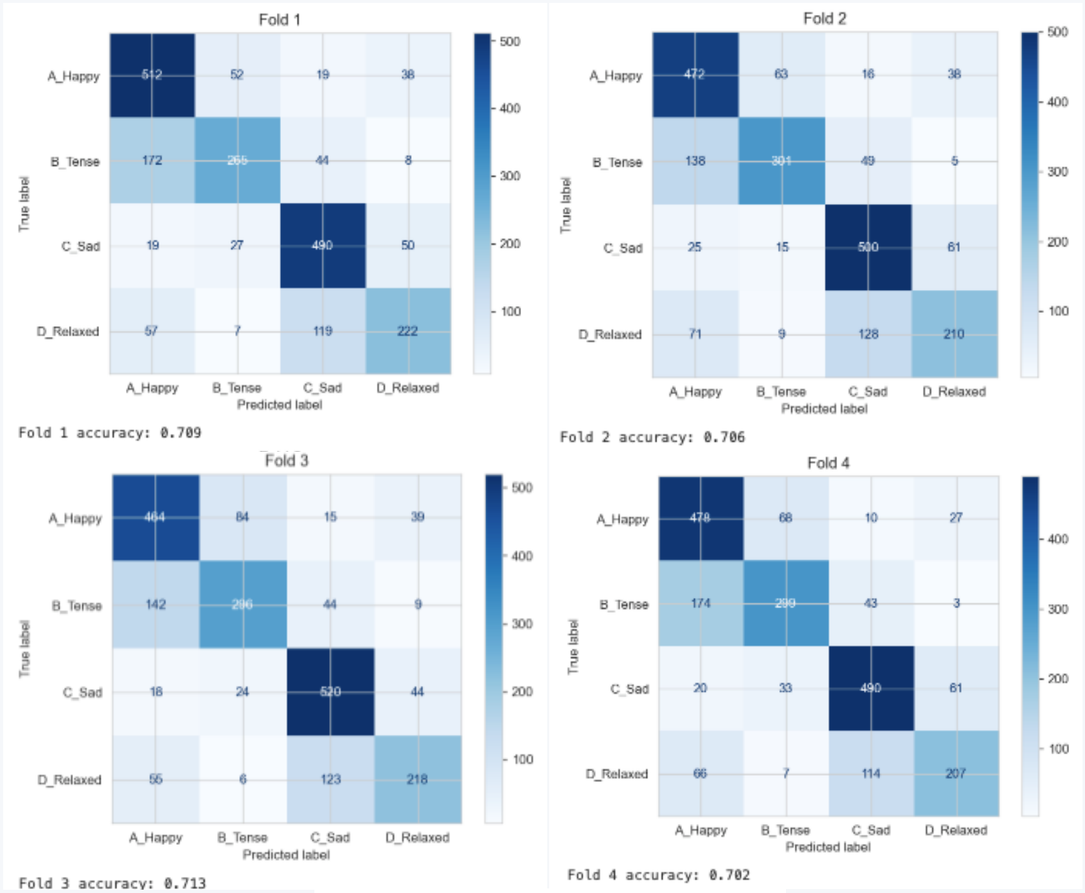
- Class imbalance was handled by applying class weighting:
- Balances the influence of each class so the model doesn't favor more frequent moods.

Making Predictions

- Each decision tree independently makes a prediction on each track.
- The final mood classification is determined by a majority vote from all trees.

Evaluating Performance

- Model performance assessed using:
- Accuracy: Correct predictions vs total predictions.
- Precision/Recall: How well each emotion is predicted specifically.
- Confusion Matrix: Shows where predictions succeed or fail (int



Results and Evalution

Cross-validation Accuracy

- 5-fold cross-validation accuracy averaged 71%.
- Stable performance across folds indicates robust predictions.

Test Set Performance

- Achieved 74% accuracy on a held-out test set (2626 tracks)
- High precision and recall across all mood categories indicate effective handling of class imbalance.

Key Findings from Confusion Matrix

- Most prediction errors occur between emotions sharing similar energy levels:
- High-energy moods: Some confusion between Happy and Tense.
- Low-energy moods: Some confusion between Sad and Relaxed.
- Confirms that model better distinguishes emotional arousal (energy) than valence

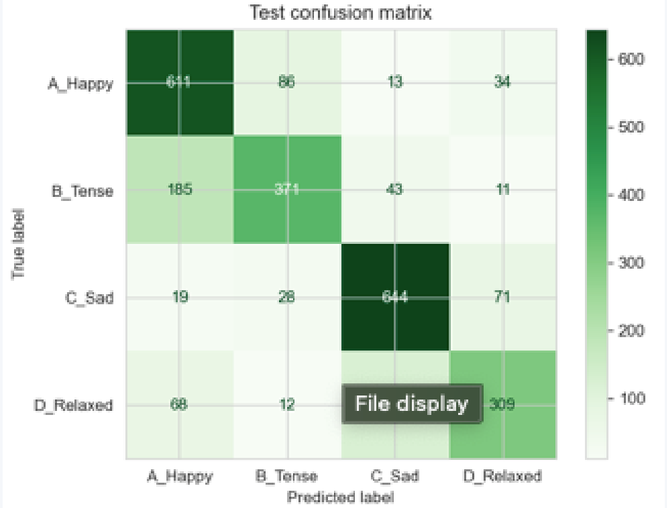
Feature Importance Insights

- Model's accuracy plateaus at about 55 beats, highlighting that the emotional signal is strongly concentrated in the first segments of a track (intro to chorus).
- Beat importance ranking provides interpretability—highlighting which parts of music are critical for emotional perception

```
best K from curve: 55
              precision    recall  f1-score   support

     A_Happy       0.69      0.82      0.75       744
     B_Tense       0.75      0.61      0.67       610
       C_Sad       0.78      0.85      0.81       762
   D_Relaxed       0.73      0.61      0.66       510

    accuracy                           0.74      2626
   macro avg       0.74      0.72      0.72      2626
weighted avg       0.74      0.74      0.73      2626
```



Conceptual Insights

Beat-level Timbre as a Strong Emotional Predictor

- Beat-level PCA timbre features contain significant emotional information, even without lyrical or metadata input.

Early Segments Matter Most

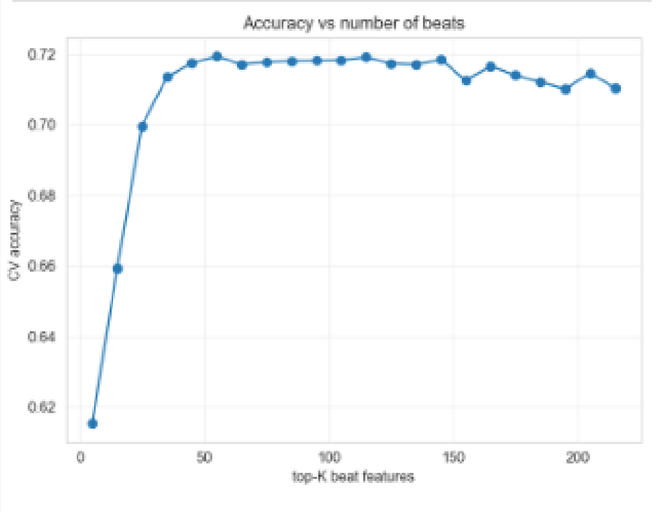
- Most emotional cues are concentrated within the first 55 beats (typically corresponding to intro and chorus sections).
- This aligns with musical intuition: intros and choruses are often the most emotionally impactful parts of songs.

Arousal vs. Valence Prediction

- The model is more accurate at distinguishing arousal (energy) levels than valence (positivity/negativity).
- Timbre (especially brightness) strongly corresponds to perceived energy/arousal, validating the psychoacoustic basis of timbre as an emotional descriptor.

Model Interpretability

- The Random Forest approach allowed for easy interpretability:
- Clear identification of which beats are crucial for emotional recognition.



Conclusions

- Successfully demonstrated that beat-level timbre data alone can predict emotional categories with high accuracy (~74%).
- Confirmed significant emotional information is concentrated in early track segments (intro and chorus).
- Established timbre features as strong indicators of emotional arousal (energy).

Recommendations for Future Work

- Explore integrating additional global features (tempo, danceability, etc.) to enhance predictive performance.
- Investigate deeper models (e.g., neural networks) to potentially improve emotion classification.
- Extend research into emotion-conditioned music generation, leveraging these predictive insights.
- Conduct listener validation studies to correlate predicted emotions with human perception explicitly.

References

FMA Dataset Reference

- Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A Dataset for Music Analysis. 18th International Society for Music Information Retrieval Conference (ISMIR 2017).

EchoNest Analyzer Documentation

- The Echo Nest Corporation. (2014). Echo Nest Analyzer Documentation v3.2.

Russell's Circumplex Model

- Russell, J. A. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

Spotify Emotion Recognition Study

- Bittner, R. M., & McFee, B. (2021). How Does the Spotify API Compare to the State-of-the-Art in Music Emotion Recognition?