Exploring Song Segmentation for Music Emotion Variation Detection

Tomás Ferreira
Hugo Redinha
Pedro Louro
Ricardo Malheiro
Rui Pedro Paiva
Renato Panda

Fundação para a Ciência e a Tecnologia

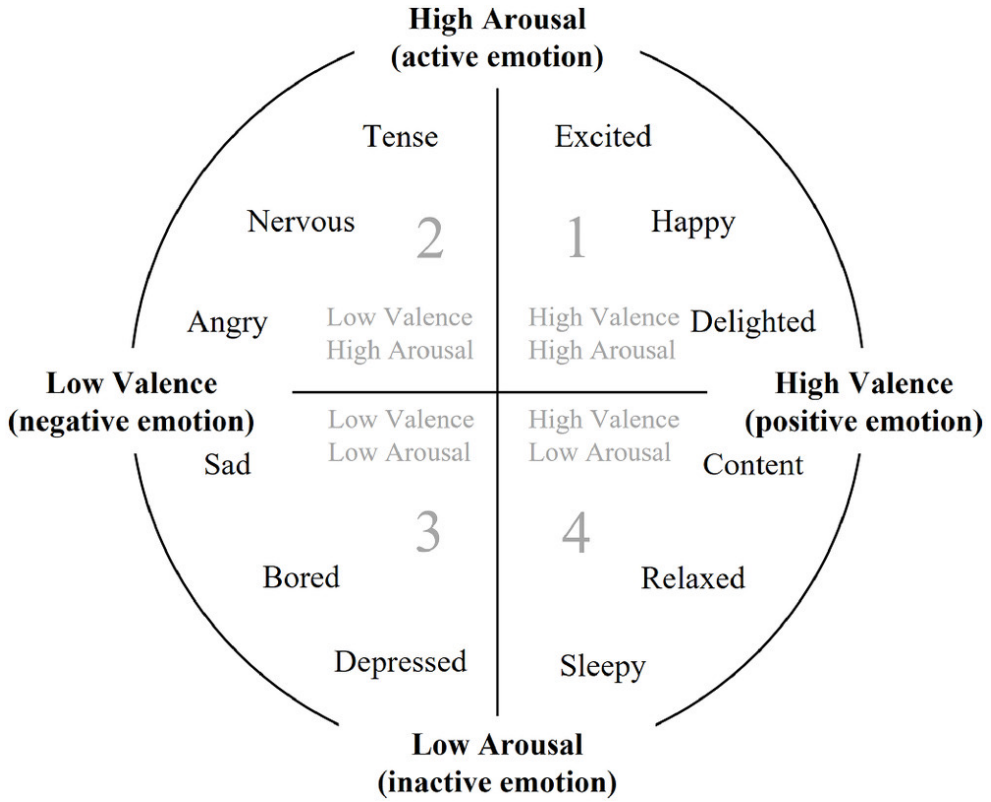
CISUC

Introduction to Music Emotion Recognition (MER) & Music Emotion Variation Detection (MEVD)

- Music Emotion Recognition (MER) involves identifying emotions conveyed by music, such as happiness or sadness.
- Music Emotion Variation Detection (MEVD) focuses on detecting changes in emotional expression throughout a song, addressing the variation of emotions over time, rather than static emotion identification.

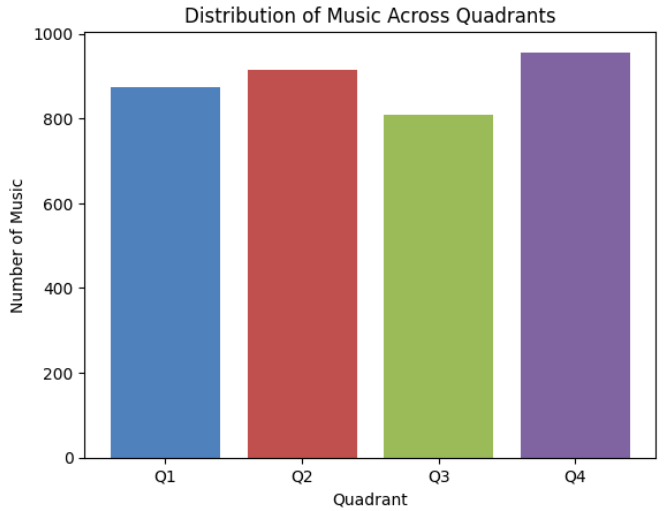
Russell's Circumplex Model

- Russell's model defines emotions based on two axes: Valence (positive to negative) and Arousal (high to low energy).
- Hence, four quadrants result:
- Q1, with high valence/high arousal (positive and energetic, e.g., happiness);
- Q2, with low valence/high arousal (negative but energetic, e.g., anger);
- Q3, with low valence/low arousal (negative and low-energy, e.g., sadness);
- Q4, with high valence/low arousal (positive but subdued, e.g., calm).

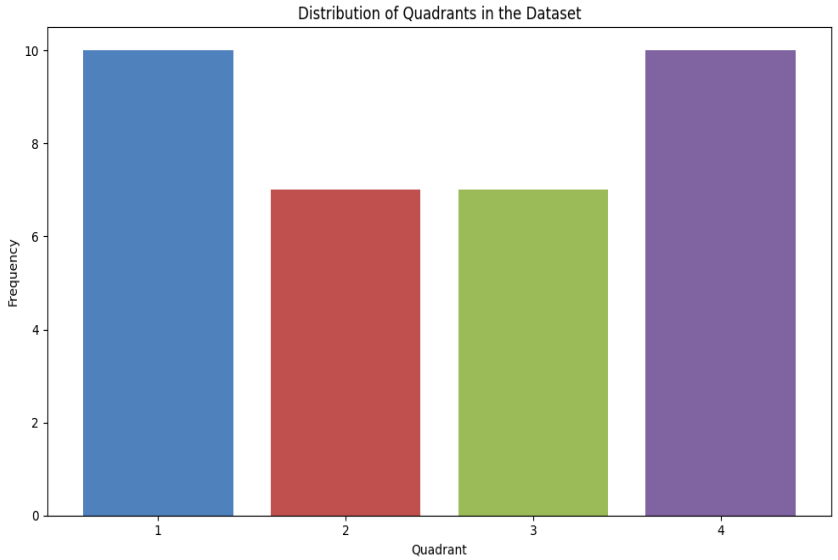


Russell's Circumplex

Dataset Overview



MERGE Audio Complete dataset song distribution across quadrants

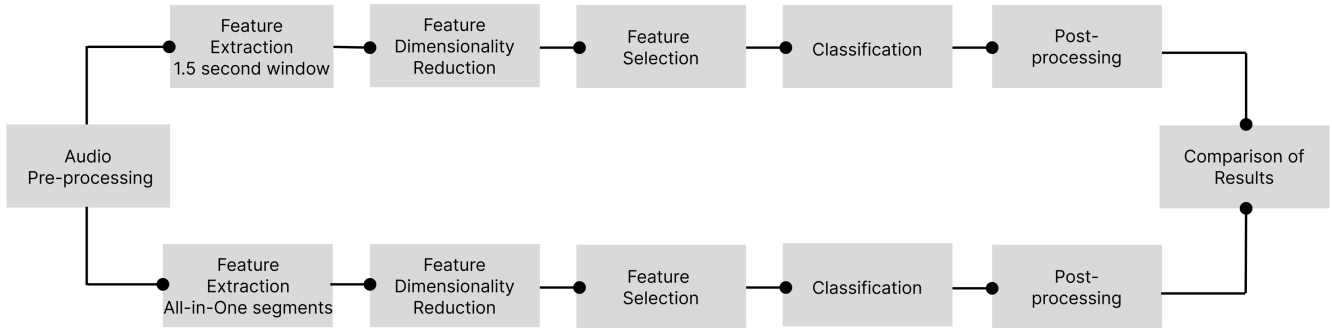


MEVD dataset song distribution across quadrants

Methodology

Two segmentation strategies:

1. Fixed-size 1.5-sec windows.
2. Variable-duration segments using the All-In-One segmentation tool



High level Overview of the Methodology

Results

	1.5 standard	All-In-One standard	1.5 novel	All-In-One novel
Q1	67.70%	35.70%	68.20%	36.20%
Q2	60.10%	19.40%	57.20%	21.40%
Q3	27.20%	23.30%	27.90%	21.00%
Q4	48.80%	29.20%	52.40%	30.20%
Weighted Avg	**53.20%**	**29.40%**	**54.80%**	**29.80%**

F1-score obtained for the 3-fold cross-validation experiment using only the 34-song dataset per quadrant.

	1.5 standard	All-In-One standard	1.5 novel	All-In-One novel
Q1	57.74%	56.98%	57.77%	56.25%
Q2	46.62%	50.46%	49.06%	44.65%
Q3	42.23%	48.00%	40.94%	44.24%
Q4	59.98%	54.72%	60.06%	50.00%
Weighted Avg	**52.97%**	**53.17%**	**53.38%**	**49.55%**

F1-score obtained with the static MER and 34-song datasets experiment per quadrant.

Discussion and Findings

- While the **All-In-One approach** generally underperformed compared to 1.5-second segments, it slightly outperformed the fixed-window approach using standard features in cross-dataset evaluation.
- The lower performance of All-In-One in the 3-fold cross-validation may be due to **fewer samples** and potential segmentation issues.
- The method segments songs into parts like intro, verse, and chorus, achieving a **70.10% F-measure** for segmentation accuracy. However, **emotional variability within segments** creates difficulties in precise emotion classification.
- The improvement seen when using the larger MERGE Audio dataset highlights the **limitations of using a small dataset**, such as the 34-song MEVD dataset, for emotion detection. More extensive and better-segmented datasets are crucial for reliable results.

Conclusions

- Song segmentation plays a significant role in MEVD.
- The All-In-One tool showed potential but didn’t achieve optimal results due to dataset size limitations.
- Improving segmentation systems like All-In-One is also a promising direction for future research.