CISUC
CENTRO DE INFORMÁTICA E SISTEMAS DA UNIVERSIDADE DE COIMBRA

IMMERSIVE
SOUND AND MUSIC COMPUTING

Exploring Song Segmentation for Music Emotion Variation Detection

Tomás Ferreira ·
tomasferreita@student.dei.uc.pt

Hugo Redinho ·
redinho@dei.uc.pt

Pedro Lima Louro ·
pedrolouro@dei.uc.pt

Ricardo Malheiro ·
rsmal@dei.uc.pt

Rui Pedro Paiva ·
ruipedro@dei.uc.pt

Renato Panda·
panda@dei.uc.pt

www.cisuc.uc.pt

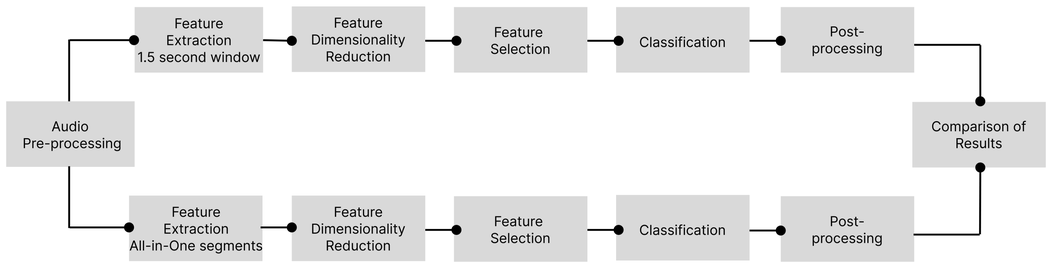

Figure 1

1. Introduction

Music Emotion Variation Detection (MEVD) focuses on tracking emotion changes within songs, addressing a gap in Music Emotion Recognition (MER) research. This study builds on the work of Panda et al.[2], to explore the impact of using fixed-size (1.5-second) and variable-size (All-in-One) window segments on MEVD performance using classical machine learning approaches. We also compare emotionally novel features (e.g., melody, harmony, rhythm) introduced by Panda et el.[1] to standard low-level spectral features.

To this end, two research questions were defined:

- How do fixed and variable window sizes affect MEVD performance in classical approaches?
- What is the impact of using emotionally relevant features compared to standard audio features?

2. Methodology

For evaluation, we created the MEVD dataset of 34 full-length songs annotated by four subjects according to Russell's quadrants. The distribution is as follows: Q1 has 10 songs, Q2 has 7, Q3 has 7, and Q4 has 10. Additionally, we used a static MER dataset, MERGE Audio Complete dataset, developed by our team, comprising 3,554 30-second song excerpts, distributed as follows: Q1 with 875 songs, Q2 with 915, Q3 with 808, and Q4 with 956.

Our evaluation involved two strategies: (i) a stratified 3-fold cross-validation with three repetitions on the 34-song dataset and (ii) training on the 3,554-song static MER dataset and testing on the 34-song dataset. All experiments were evaluated using the F1-score metric.

Figure 1 provides a high-level overview of the methodology, illustrating the process from raw audio samples to final emotion prediction.

3. Results

In our 3-fold cross-validation experiment, 1.5-second segments achieved F1-scores of 53.20% with standard features and 54.80% with novel features, indicating that the novel features slightly outperformed the standard ones. The results of this 3-fold cross-validation are presented in Table 1. However, when using the All-In-One approach, the F1-scores were significantly lower, with 29.40% for standard features and 29.80% for novel features.

In the experiment that trained on the static dataset and tested on the MEVD dataset, 1.5-second segments increased the F1-score from 52.97% to 53.17% with novel features. The All-In-One approach also showed a slight improvement over 1.5-second segments when using standard features, raising the F1-score from 52.97% to 53.17%. The results of this experiment are detailed in Table 2.

4. Discussion and Limitations

While the All-In-One approach generally performed less than the 1.5-second segments in most cases, it slightly outperformed the fixed-window approach when using standard features in the cross-dataset evaluation. The lower performance of All-In-One in the 3-fold cross-validation could be due to the smaller number of samples and potential issues with segmentation accuracy. The All-In-One method segments songs into categories such as intro, verse, chorus, solo, outro, and bridge, achieving a weighted average F-measure of 70.10% for segmentation accuracy on the MEVD dataset.

	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%**

Table 1

	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%**

Table 2

While this segmentation is promising, more is needed for precise emotion-based analysis, as All-In-One segments could contain multiple emotions, as shown in Figure 2. In this figure, we can see the segment from 97 to 124 seconds, divided into two parts: 97-100 seconds (Q1) and 100-124 seconds (Q4), indicating the presence of multiple emotional shifts within the segment. This lack of emotional uniformity within segments poses difficulties to the classification process, which might partly justify the low scores attained. The improvement seen when using the larger MERGE Audio Complete dataset for training and testing on the 34-song MEVD dataset underscores the limitations of using a small dataset, like the 34-song MEVD dataset, in the 3-fold cross-validation experiment. These findings highlight the importance of having a more extensive training dataset and high-quality segmentation for reliable emotion detection.

5. Conclusion

This work studied the impact of song segmentation in MEVD. In particular, All-In-One was employed but did not completely fulfill its promise for MEVD despite its potential and encouraging preliminary results. As discussed, this might be a consequence of the small employed dataset. Therefore, we are now working on the creation of a larger MEVD dataset, keeping the objective of having quality annotations. Another promising line of future research regards the improvement of song-structure segmentation systems such as All-In-One.

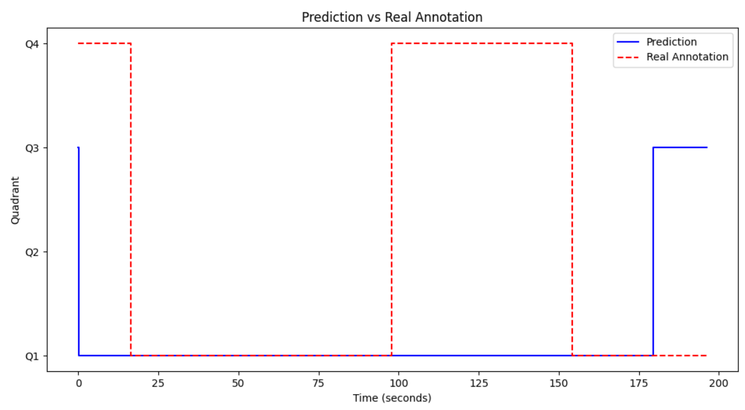

Figure 2

References

[1] Panda, R., Malheiro, R., and Paiva, R. P. Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing 11*, 4 (2020), 614–626.

[2] Panda, R., and Paiva, R. P. Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In *130th Audio Engineering Society Convention 2011 (AES)* (London, UK, 2011), pp. 579–586.

Captions of Figures

Figure 1: A high-level overview of the methodology for detecting music emotion variations by comparing fixed-size 1.5-second windows and variable-size all-in-one segments. Following audio pre-processing, both workflows undergo feature extraction, dimensionality reduction, feature selection, classification, and post-processing. The results are then compared to assess the effectiveness of each window size in capturing emotion variations.

Figure 2: Comparison between the annotated and predicted emotion quadrants for the song "Whenever, Wherever" by Shakira using the All-in-One segmentation approach.

Captions of Tables

Table 1: F1-score obtained for the 3-fold cross-validation experiment using only the 34-song dataset per quadrant.

Table 2: F1-score obtained with the static MER and 34-song datasets experiment per quadrant.

UNIVERSIDADE D COIMBRA 1290

FCT Fundação para a Ciência e a Tecnologia

ERGE Music Emotion Recognition Next Generation