Exploring Song Segmentation for Music Emotion Variation Detection

Tomás G. Ferreira[1] tomasferreira@student.dei.uc.pt[0009−0006−5102−6915],
Hugo Redinho[1] redinho@dei.uc.pt[0009−0004−1547−2251],
Pedro L. Louro[1] pedrolouro@dei.uc.pt[0000−0003−3201−6990],
Ricardo Malheiro[1,2] rsmal@dei.uc.pt[0000−0002−3010−2732],
Rui Pedro Paiva[1] ruipedro@dei.uc.pt[0000−0003−3215−3960], and
Renato Panda[1,3] panda@dei.uc.pt[0000−0003−2539−5590]

[1] University of Coimbra, CISUC, DEI, LASI, Portugal
[2] School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[3] Ci2 — Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal

Abstract. This paper evaluates the impact of song segmentation on Music Emotion Variation Detection (MEVD). In particular, the All-In-One song-structure segmentation system was employed to this end and compared to a fixed 1.5-sec window approach. Acoustic features were extracted for each obtained segment/window, which were classified with SVMs. The attained results (best F1-score of 55.9%) suggest that, despite its promise, the potential of this song segmentation approach was not fully exploited, possibly due to the small employed dataset. Nevertheless, preliminary results are encouraging.

Keywords: MEVD · Song Segmentation

1 Introduction

Music Emotion Recognition (MER) aims to identify emotions conveyed by music, such as happiness or sadness. Some of the most critical MER challenges include the need for accurate feature extraction, quality and sizeable datasets, and the subjective nature of human annotation.

Most research works tackled static MER, i.e., the identification of the dominant single emotion in a song [6]. Besides static MER, Music Emotion Variation Detection (MEVD) is an important topic, as it is well known that emotions vary throughout songs [1]. However, it has received less research attention.

Two strategies can be followed to tackle MEVD: i) analyze short fixed-length segments [7]; ii) perform emotion-based segmentation and analyze the obtained segments [4]. In the first, more straightforward approach, MEVD can be regarded as a direct extension of static MER, where the employed features and classifiers can be used to classify short windows (typically up to 2-sec). In the second strategy (rarely applied), emotion-based segmentation is performed beforehand, after which the obtained segments are classified.

Regarding the second approach, music segmentation tools like All-In-One [3] have potential. These tools create windows with dynamic lengths that match the natural segments of a song, such as the chorus, verse, and bridge. The usage of the All-In-One model is due to its ability to perform structural segmentation, which is crucial for accurate emotion analysis within music. The All-In-One model is a deep learning-based tool for comprehensive music structure analysis. It performs tasks such as beat and downbeat tracking, segmentation of a song into distinct sections, and labelling these segments based on their roles (e.g., verses, choruses). Leveraging advanced transformer architectures with dilated neighbourhood attention mechanisms, the model captures both local details and long-term patterns, enabling accurate segmentation and precise labelling. This ensures that the model detects where one section ends and another begins and understands the context of each segment within the song's overall structure.

One frequently used emotion taxonomy is Russell's Circumplex Model [8], which characterizes emotion based on two dimensions: 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); and Q4, with high valence/low arousal (positive but subdued, e.g., calm).

Several works can be found in the MEVD literature, either following classical feature engineering with machine learning (ML) [7] and deep learning (DL) [5] approaches. In the first group, handcrafted features, namely, music-content features capturing concepts such as tempo, key, or low-level spectral descriptors, have been devised [6], and used as input to classical ML classifiers, e.g., Support Vector Machines (SVM). One example in this category is the work by Panda et al. [7], where the authors proposed a system based on SVMs to predict changes in emotion quadrants using audio features on fixed-size windows. The system analyzed 189 clips from various genres, achieving an average F1-score of 53.71%, with higher accuracy for positive valence predictions.

As for DL, Recurrent Neural Networks (RNN) and Long-Short-Term-Memory (LSTM) Networks offer interesting possibilities for handling temporal information by tackling the long-term dependency problem. However, they lack the adaptative learning power of Convolutional Neural Networks (CNN). Thus, some works combine CNNs and RNN/LSTMs [5].

Regardless of the employed approaches, current MEVD results are still underperforming, e.g., in the DEAM dataset created for the MediaEval challenge [1], the highest attained correlation coefficient scores were 0.19 and 0.52 for valence and arousal, respectively. As in static MER, the underperformance in MEVD stems from the lack of emotionally relevant features and quality and sizeable datasets [6].

In this article, we build on Panda et al. [7] to tackle the following goals:

1. To compare both window- and segmentation-based approaches. Here, we aim to employ the All-In-One segmentation tool.

2. To evaluate the impact of employing the set of emotionally relevant features proposed by Panda et al. (including musical dimensions such as melody, harmony, rhythm, dynamics, expressivity, and texture) in comparison to standard audio features, primarily based on low-level spectral features [6].
3. To evaluate the previous models on a small yet controlled MEVD dataset created by our team containing 34 full-length songs.

The best-performing model employing song segmentation attained an F1-score of 53.17%, which was below the top score using 1.5-sec windows (54.8%). This suggests its potential was not fully exploited, as will be discussed.

2 Materials and Methods

As mentioned above, we build on [7] to compare the influence of different segmentation strategies in MEVD: variable segments from All-in-One and fixed-size 1.5-sec segments. We also examine the impact of standard and novel [6] emotion-related features. To this end, we evaluate the proposed methods on a 34-song dataset.

2.1 Datasets and Evaluation Strategy

For evaluation purposes, we created a dataset containing 34 full-length songs. Four subjects annotated the songs according to Russell's quadrants. The distribution of the songs per quadrant is as follows (regarding the prevalent quadrant): Q1 has 10 songs; Q2 has 7; Q3 also has 7; and Q4 has 10 songs.

Moreover, we also employed a static MER dataset created by our team (yet to be published) containing 3554 30-second song excerpts. These songs were distributed across Russell's quadrants: Q1 with 875 songs; Q2 with 915; Q3 with 808; and Q4 with 956. Emotion annotations are determined by an absolute majority consensus among multiple annotators, ensuring collective agreement and enhancing reliability.

All audio clips (in the two datasets) were standardized to WAV PCM format with a sampling rate of 22050 Hz, 16-bit quantization, and mono-aural configuration.

Both datasets were segmented using fixed 1.5-sec windows and variable-duration segments with the All-In-One tool [3]. The choice of window size for extracting emotional features is particularly critical. One research [9] showed that it takes different durations to recognize various emotions: happy (483 ms), sad (1446 ms), scary (1737 ms), and peaceful (1261 ms). Emotions like scary and peaceful take longer to detect due to their complexity and lower energy levels. In [7], several experiments were conducted, testing fixed 1.5 sec windows and the variable segments produced by the All-In-One tool, where best results were attained with 1.5-sec windows. This value corresponds with the mentioned durations, so we kept it in this article.

In terms of evaluation, we followed two strategies: i) stratified 3-fold cross-validation, with thirty repetitions (in the 34-song dataset); ii) training on the 3554-song static MER dataset and testing in the 34-song dataset.

Given the small dataset size, we chose a low number of folds (only three) for the first strategy. To prevent data leakage, we guarantee that different segments from the same song cannot appear simultaneously in the training and test folds (i.e., all segments of a song must be placed either in the training or test set to avoid overly optimistic results).

The second strategy aims to assess the impact of training our model on a larger dataset.

Finally, all experiments are evaluated with the F1-score metric.

2.2 Feature Engineering and Classification

Two sets of features previously introduced in [6] were used, hereafter referred to as standard and novel features.

The standard features, totaling 1604, were derived from the MIR Toolbox, Marsyas, and PsySound audio frameworks and encompass dimensions such as melody, rhythm, harmony, and dynamics, with the majority being tone color. These features include MFCCs, zero-crossing rate, spectral centroid, spectral flux, and chroma features. Given the overlap of these frameworks, features with high correlation and low relevance were removed, as described in [6].

The novel features [6] capture higher-level musical concepts previously ignored by standard features. They are primarily derived from the notes of the melody, recreated through the estimation of fundamental frequencies and pitch saliences, which are then segmented into individual MIDI notes. These features enable extracting high-level features related to melody, dynamics, rhythm, musical texture, expressivity, and tone color. They were also extracted from the isolated singing voice using source separation tools, acknowledging the significance of vocal cues.

After extracting the features, we performed dimensionality reduction. Initially, we removed features with zero standard deviation. Then, we eliminated heavily correlated features with a correlation factor exceeding 0.9. Next, we used the RelieF algorithm to select the most important features by ranking them based on their importance. We analyzed various subsets of these top-ranked features, with subset sizes ranging from 5 to 1250. Ultimately, the model demonstrated optimal performance with 900 features.

Finally, SVM models were trained with the described datasets. Here, we used Bayesian search [2] to find the best hyperparameters for our SVM models. This involved adjusting kernel type, gamma, cost, and polynomial degree. We considered kernel types such as Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. The cost was searched from 1e-6 to 100, while gamma ranged from 1e-6 to 5000. The polynomial degree ranged from 1 to 5 (where applicable).

2.3 Post-Processing

The experiment with the static windows of 1.5 seconds used a median filter to smooth the data and reduce outliers. This filter works by sliding a window over the dataset and replacing each data point with the median of its neighbors. This effectively removes sharp spikes while maintaining the overall shape and trends of the data. A specialized filter was used for the All-In-One segments to account for varying segment sizes. This filter evaluated the length of each segment and checked if shorter segments (less than one second) shared the same quadrant as their neighboring segments. If so, the current segment's quadrant value was updated accordingly, ensuring continuity and coherence.

3 Results and Discussion

Table 1 and Table 2 provide a comprehensive summary of the results for the 3-fold experiment and the experiment using the two datasets, respectively.

Table 1. F1-score obtained for the 30x3-fold CV experiment using only the 34-song dataset per quadrant.All numbers are expressed as percentages.

	1.5 standard	All-in-One standard	1.5 novel	All-in-One novel
Q1	68.9	36.3	67.8	37.1
Q2	62.4	24.5	62.9	27.3
Q3	24.6	19.5	25.4	19.6
Q4	51.2	26.7	53.9	28.4
Weighted Avg	**55.1**	**29.9**	**55.9**	**30.6**

Table 2. F1-score obtained with the static MER and 34-song datasets experiment per quadrant.All numbers are expressed as percentages.

	1.5 standard	All-In-One standard	1.5 novel	All-In-One novel
Q1	57.7	57.0	57.8	56.3
Q2	46.6	50.5	49.1	44.7
Q3	42.2	48.0	41.0	44.2
Q4	60.0	54.7	60.0	50.0
Weighted Avg	**53.0**	**53.2**	**53.4**	**49.6**

In our 30x3-fold CV experiment, we observed that the 1.5-second segments achieved F1-scores of 55.1% with standard features and 55.9% with novel features, indicating that the novel features slightly outperformed the standard ones. However, the results suggest room for improvement and that the small dataset size may have affected the lower classification performance.

Conversely, when employing the All-in-One approach, we observed significantly lower scores of 29.9% and 30.6% for standard and novel features, respectively.

The statistical tests confirm these observations:

1. For 1.5-second segments, the comparison between novel and standard features showed no significant difference (p-value = 0.50061), indicating that while novel features performed slightly better, the improvement was not statistically significant.
2. Similarly, for the All-in-One approach, the comparison between novel and standard features also showed no significant difference (p-value = 0.55895).
3. However, when comparing the 1.5-second segments to the All-in-One approach, both for novel features (p-value = 7.1982×10^{-55}) and standard features (p-value = 2.5008×10^{-51}), the differences were statistically significant. This indicates a clear difference in performance based on the segmentation method used.

Interestingly, when using novel features for All-in-One segments, the top-ranked features were mainly standard features. Only a few new features made it to the top, justifying the small improvement from standard to novel features and suggesting that most novel features do not improve the outcomes for All-in-One segments. The increased complexity of new features comes from their specificity in capturing a single emotion. If a segment contains multiple emotions, these features provide little to no valuable information, resulting in poorer performance. Essentially, these features are more complex than standard ones and are effective only when a segment has a single emotional tone. When multiple emotions are present, the effectiveness of these new features decreases significantly.

The lower results could be attributed to the small dataset size, which may have contributed to challenges in the classification process. Moreover, the obtained segments may each contain multiple emotions. The absence of emotional consistency within the segments makes it challenging to classify them, possibly contributing to the low score achieved. Figure 1 illustrates the predicted and real emotional variation for one song.

In this particular song, there exists a segment that spans from 97 seconds to 124 seconds. According to the annotation, the segment is divided into two parts: from 97 seconds to 100 seconds, characterized as Q1, and from 100 seconds to 124 seconds, characterized as Q4. This suggests that the segment in question may potentially encompass multiple emotional shifts.

Another cause might be the segmentation outcomes. All-In-One segments songs into six segment types: intro, verse, chorus, solo, outro, and bridge. In our 34-song dataset, this segmentation achieved a weighted average F-measure of 70.10%. Although this is a promising performance (above 66% reported in the article [3]), it might be insufficient for accurate emotion-based segmentation. Hence, each of the obtained segments might contain more than one emotion. This lack of emotional uniformity within segments poses difficulties to the classification process, which might partly justify the low attained score.

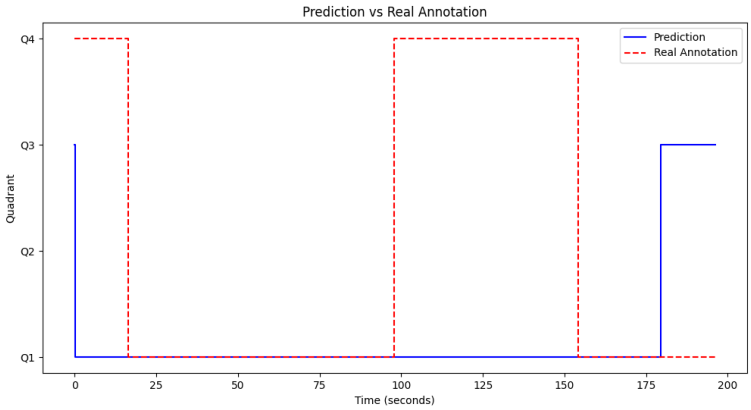

Fig. 1. Comparison between the annotated and predicted emotion quadrants for the song "Whenever, Wherever" by Shakira using the All-in-One segmentation approach.

In the experiment using the static dataset for training and the 34-song dataset for testing, 1.5-sec window size increased the F1-score from 53.0% to 53.4% with the novel features. Moreover, All-In-One slightly improved over 1.5-sec indow size, from 53.0.% to 53.2% with the standard features. However, the novel features did not perform as well as the standard features when using the All-in-One approach, decreasing from 53.2% to 49.6%. The decrease in performance for the All-in-One segments is attributed to the fact that feature ranking was computed for the MERGE Audio Complete dataset, not the MEVD dataset, causing many novel features to rank highly. As seen in the 30x3-fold CV cross-validation experiment, the All-in-One segments performed poorly with most novel features. These features are more intricate than standard ones and are only adequate for segments with a single emotional tone. All-in-One segments often contain multiple emotions, as shown in Figure 1, causing top features to offer little information, resulting in poorer outcomes.

4 Conclusions and Future Work

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.

Acknowledgments

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - DOI: 10.54499/PTDC/CCI-COM/3171/2021 financed with national funds (PIDDAC) via the Portuguese State Budget; and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.

References

1. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PLoS ONE **12**(3) (2017). https://doi.org/10.1371/journal.pone.0173392
2. Brownlee, J.: Probability for Machine Learning: Discover How To Harness Uncertainty With Python. Machine Learning Mastery (2019), https://books.google.pt/books?id=uU2xDwAAQBAJ
3. Kim, T., Nam, J.: All-In-One Metrical And Functional Structure Analysis With Neighborhood Attentions on Demixed Audio. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY (2023)
4. Lu, L., Liu, D., Zhang, H.J.: Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Transactions on Audio, Speech, and Language Processing **14**(1), 5–18 (2006). https://doi.org/10.1109/TSA.2005.860344
5. Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., Jarina, R.: Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. In: 14th Sound & Music Computing Conference (SMC). pp. 208–213. Espoo, Finland (2017)
6. Panda, R., Malheiro, R., Paiva, R.P.: Novel Audio Features for Music Emotion Recognition. IEEE Transactions on Affective Computing **11**(4), 614–626 (2020). https://doi.org/10.1109/TAFFC.2018.2820691
7. Panda, R., Paiva, R.P.: Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In: 130th Audio Engineering Society Convention 2011 (AES). pp. 579–586. London, UK (2011)
8. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980). https://doi.org/10.1037/h0077714
9. Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., Bouchard, B.: Happy, sad, scary and peaceful musical excerpts for research on emotions. Cognition & Emotion **22**(4), 720–752 (2008). https://doi.org/10.1080/02699930701503567