Exploiting Automatic Source Separation and Music Transcription for Music Emotion Recognition

Hugo Redinho[1], Pedro Lima Louro[1], André C. Santos[1], Ricardo Malheiro[1,2], Rui Pedro Paiva[1], and Renato Panda[1,3(✉)]

[1] University Coimbra, CISUC, DEI, LASI, Coimbra, Portugal
redinho@student.dei.uc.pt,
{pedrolouro,andresantos,rsmal,ruipedro,panda}@dei.uc.pt
[2] Polytechnic Institute of Leiria, School of Technology and Management, Leiria, Portugal
[3] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Tomar, Portugal

Abstract. We present a set of novel emotionally relevant audio features related to percussion and individual instrument information to help improve the classification of emotions in music. To this end, as intermediate steps, we performed automatic music transcription (using the MT3 framework) and music source separation (using the Demucs framework). Leveraging the outcomes of these two frameworks, we developed a set of novel features that capture information related to musical texture, rhythm, and melody, which are among the least represented musical dimensions in Music Emotion Recognition (MER). To validate our work, we employed the recently created MERGE dataset, which contains over 3000 30-s audio clips annotated in terms of Russell's emotion quadrants. To assess the impact of the proposed features, we compared the classification results obtained in this dataset with current state-of-the-art features. The conducted experiments show that the novel features improved the music emotion classification results. Moreover, the best-performing approach achieved an F1-score of 74.1% and employed 200 features (after feature selection), of which 62 were novel.

Keywords: Music Emotion Recognition · Music Information Retrieval · Feature Engineering · Machine Learning

1 Introduction

With the massive growth of available music in recent years, there have been increasing efforts to categorize and enable filtering of this vast amount of data to make it possible to cater to every person's interests. The field of Music Emotion Recognition (MER) aims to tackle this problem by creating tools involving machine learning (ML) techniques to identify the emotions present in songs.

S. Palaiahnakote et al. (Eds.): ICPR 2024, LNCS 15614, pp. 67–82, 2025.
https://doi.org/10.1007/978-3-031-87657-8_5

There have been many different problems addressed in this realm, for example, music emotion classification [7,14,31], music emotion variation detection [17,24], automatic playlist generation [8,19], amongst others.

However, due to the nature of music, it is hard to categorize songs, partly due to the sheer amount of variables that can change from song to song, such as singer, the number of instruments, tonality, melody, tempo, and dynamics, amongst many other characteristics.

Furthermore, in the audio MER realm, two additional problems complicate MER tasks: 1) the lack of emotionally relevant audio features, with most approaches using a similar set of features that were originally proposed to address other audio analysis problems (e.g., speech recognition), and lacking emotional relevance [23]; and 2) the lack of sizeable and quality datasets, which further limits the impact of techniques such as deep learning that require a large amount of data. Moreover, it also poses difficulties in benchmarking different MER approaches.

In a recent survey by our team [23], we highlighted the potential impact of developing novel, emotionally relevant acoustic features. e also demonstrated that features suited explicitly for emotion detection are needed to narrow the so-called semantic gap and break the current glass ceiling in MER (as well as in other Music Information Retrieval (MIR) problems) [3].

As such, in this work, we propose a set of novel features that take advantage of automatic music transcription using Magenta MT3 [11] and music source separation using Demucs [27]. Leveraging the output of these two frameworks and building on our previous work [22], we developed a set of novel features that capture information related to musical texture, rhythm, and melody, which are among the least represented musical dimensions in MER [23].

To validate our work, we used the MERGE [16] dataset containing over 3000 30-s song excerpts, briefly described in the following sections.

To evaluate the effectiveness of the proposed features, we conducted a comparison of the classification results from this dataset with the current state-of-the-art features from Panda et al. [22] (hereafter termed "baseline features"), using Support Vector Machine (SVM) classifiers and the ReliefF feature selection algorithm.

Our experiments show that, in the best-performing model, the novel features improved the music emotion classification results from 71.0% to 74.1% compared to the baseline features. Furthermore, this model employed 200 features (after feature selection), out of which 62 were novel. Moreover, the novel features helped to reduce the confusion between quadrants 3 and 4 (a typical problem in MER classification).

Despite the potential of deep learning and bimodal solutions (which our team has also addressed, e.g., in [16]), the aim of this paper is to focus the analysis on the impact of new emotionally relevant acoustic features, given its key role in MER.

The paper is organized as follows: Sect. 2 reviews the related work. Section 3 describes the proposed methods, including datasets and features (baseline and

novel). Section 4 discusses the conducted experiments, the attained results and their implications. Finally, Sect. 5 concludes with a summary of findings and future research directions.

2 Related Work

Grasping the concept of emotion has always been a challenge for humans. There have been many attempts at defining emotion, as well as dividing its various types. In the MER field, emotions are usually divided into three levels: expressed, perceived, and felt (also called induced) emotions [10]. Similarly to other works in the literature, this work is focused on the perceived emotion.

Psychological researchers have long debated the representation and classification of emotions, resulting in the introduction of various emotion paradigms (such as categorical or dimensional) and their respective taxonomies (e.g., Hevner [12] or Russell's [28] emotion models).

In this work, we employ Russell's circumplex model of emotion. This model has been supported by several studies [25], and it is also the model used in our previous work (Panda et al. [22]). This is a two-dimensional model based on two main axes: valence (pleasure-displeasure, i.e., the polarity of emotion in terms of positive and negative states, also known as pleasantness) and arousal (aroused-not aroused, also known as activity or energy). The result, represented in Fig. 1, is a two-dimensional plane (arousal-valence) where the X-axis represents valence and the Y-axis represents arousal.

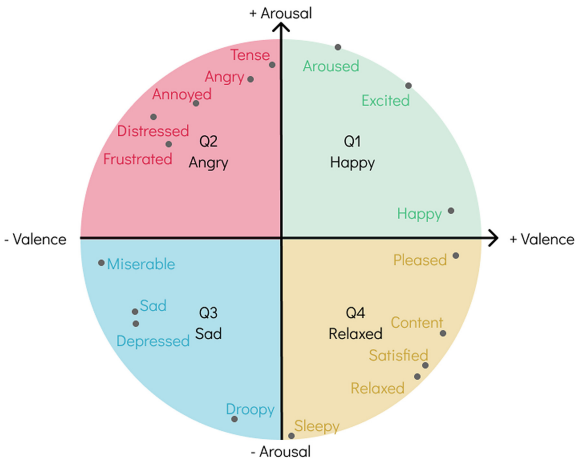

Fig. 1. Russell's Circumplex Model of Emotion (adapted from [28]).

The resulting framework consists of four quadrants characterized as follows: 1) Quadrant 1 (Q1) corresponds to positive valence and arousal, representing

happy and energetic emotions like excitement or enthusiasm; 2) Quadrant 2 (Q2) reflects negative valence and positive arousal, encompassing frantic and energetic emotions such as anxiety, fear, or anger; 3) Quadrant 3 (Q3) entails negative valence and arousal, portraying melancholic and sad emotions like depression; 4) Quadrant 4 (Q4) denotes positive valence and negative arousal, indicative of calm and positive emotions like contentment or serenity.

Regarding the relations between musical attributes and emotion, many features, such as articulation, dynamics, harmony, loudness, musical form, pitch, rhythm, amongst others, have been previously linked to emotion [9,15,19]. Many of these relations are yet to be fully understood and require further research. Moreover, there are certain attributes that are harder to extract and quantify from audio signals.

These musical attributes can be organized into eight categories [21], each representing a musical concept: melody, harmony, rhythm, dynamics, tone color (or timbre), expressivity, musical texture, and musical form.

In Panda et al. [23], our team performed a thorough review of the current status of feature engineering in the MER field. This work showcased that, of the eight musical dimensions, several are still underrepresented in MER (and MIR), namely texture, melody, expressivity, and rhythm. Moreover, it has been shown that the underrepresented musical dimensions, particularly texture, are useful, especially in identifying songs pertaining to the first and second quadrants [22]. Thus, creating new features that help represent these musical dimensions is important.

To advance the creation of novel emotionally relevant features, in this work, we resort to two frameworks: Magenta MT3 and Demucs. Magenta MT3 [11] is an automatic music transcription system that also assigns notes to their respective instruments (a maximum of 128, according to the General MIDI Level 1 standard). The output of MT3 is a MIDI file with the resulting music transcription. We leverage this framework's outcome to create melody and texture features. Moreover, Demucs [27] performs music source separation, splitting the original audio track into its composing stems (e.g., vocals, drums). Furthermore, this framework achieves the best results when compared to the music source separation state of the art. We use the output of this framework to create features related to rhythm, particularly features related to the percussive content.

3 Materials and Methods

In this section, we start by briefly introducing the datasets used in this work. Next, we briefly describe the employed baseline features. We then detail the novel features proposed in this work, organized into the different musical dimensions. Finally, we detail the evaluated emotion classification strategies.

3.1 Dataset

In this work, we employed the MERGE dataset [16] recently created by our team, with more than 3000 30-s audio clips annotated into Russell's emotion quadrants. Below, we briefly describe the employed dataset. Full details are provided in [16].

Six subsets were created from the full set of annotated songs, as summarized in Table 1. These datasets are referred to as MERGE followed by the type ("audio", "lyrics", and "bimodal") and by the balancing condition ("complete" if it is the complete and unbalanced data, or "balanced" if it is a balanced subset of the complete data, where each quadrant has the same number of songs.).

Table 1 shows the total number of songs for each of the six aforementioned datasets and the number of songs per quadrant.

Table 1. Number of songs in the MERGE datasets.

Dataset Name	Q1	Q2	Q3	Q4	Total
MERGE Audio Complete	875	915	808	956	3554
MERGE Audio Balanced	808	808	808	808	3232
MERGE Lyrics Complete	600	710	621	637	2568
MERGE Lyrics Balanced	600	600	600	600	2400
MERGE Bimodal Complete	525	673	500	518	2216
MERGE Bimodal Balanced	500	500	500	500	2000

This dataset contains a lyrics component, which is not addressed in this work. As such, the only datasets employed in the present article are the ones containing audio clips, i.e., the Audio and Bimodal datasets.

3.2 Baseline Features

As previously mentioned, we build on the set of features proposed by our team in [22], which we denote "baseline features". These features cover the usual eight musical dimensions and are split into two categories: 1) standard features, which are features extracted from audio frameworks, such as Marsyas [30], MIR Toolbox [13], and PsySound3 [2], and employed in most works prior to Panda et al.; and 2) new features, i.e., the set of features proposed by Panda et al., which are related to higher-level musical concepts, namely expressivity, and texture.

Initially, the standard set contained 1603 features, and the novel feature set contained 1086. Of these 1086 features, half (543) were obtained from the original sound waveform, and the other half resulted from the vocal-separated audio clip. After feature redundancy analysis [22], the baseline feature set comprises 1255 acoustic features: 898 from the standard set and 357 from the new set. Figure 2 [22] illustrates the baseline feature distribution across the eight musical dimensions. Further details about these baseline features are available in [22].

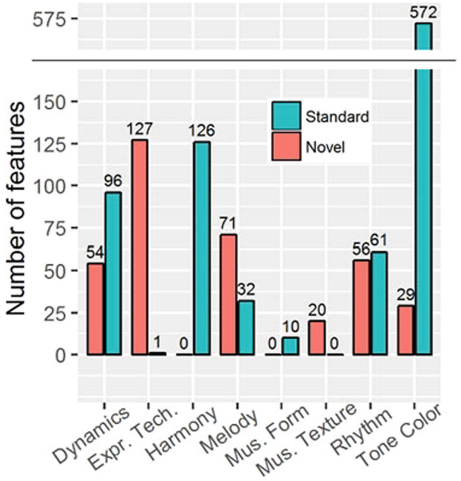

Fig. 2. Distribution of features for each musical dimension, separated by feature type (adapted from [22]).

3.3 Novel Features

As aforementioned, this work proposes a set of novel features by taking advantage of automatic music transcription and music source separation. In the next section, we describe the proposed features, which are organized into the respective musical dimensions.

Preliminary Steps. The songs are converted into a standardized audio format using the FFmpeg framework, with the following specifications: WAV PCM Format, 22050 HZ sampling rate, 16-bit quantization, and mono-aural.

After this standardization is complete, we use Demucs to perform two audio separations: 1) separate the vocal track and non-vocal track, and 2) separate the drums part and the non-drums (melodic) part of the track. These audio stems are then standardized to the aforementioned formats.

Using MT3, MIDI files with the notes for each instrument were extracted. As MT3 was trained on files with no vocals, the track with no vocals (resulting from the Demucs output) was used as input for MT3. This complements the analysis of the main melodic line, conducted mostly in the vocals, as performed in Panda et al. [22].

Melody and Texture Features. Regarding melodic and texture features, we employ MT3 [11] as a basis to: 1) update and improve previous features extracted in Panda et al. [22]; 2) create novel features.

In Panda et al. [22], the extraction of melodic features relied on the melody detection algorithm proposed by Dressler [5]. As such, only features from the main melody were created. To improve this approach, in this work, we exploit

MT3 to perform full music transcription (using the non-vocal stem, as aforementioned). This enables the analysis of all the melodic lines in a song. Therefore, the melodic features proposed in [22] are now extracted from all the melodic lines instead of only the main melody.

As for texture features (mostly related to the musical layers present, i.e., the number of simultaneous notes at each moment), Panda et al. [22] relied on the multi-pitch estimator by Dressler [5] (an extended version of Dressler's melody detection algorithm, from which multiple pitches in each short-time frame were estimated). Here, we extract the same musical texture features but now resort to MT3, as a more accurate approach, where the number of simultaneous notes is obtained.

The above features are not strictly novel, in the sense that the same algorithms are applied, although to the more accurate information provided by MT3.

Hence, besides these updated features, we propose several novel features, as described below. To this end, it was important to classify the 128 MIDI instruments into groups. The classification most commonly used in the literature is the one proposed by Sachs et al. [29]. Therefore, that was the one chosen for our work. The authors organize musical instruments into five groups: idiophones, chordophones, membranophones, aerophones, and electrophones. Moreover, the MIDI standard also performs a categorization by grouping instruments into "melodic" or "percussive". We also employ this binary organization in this work.

Therefore, we propose the following novel features:

1. **Instrument extent.** Before, measuring the presence or absence of a certain instrument in an audio track was impossible. With the transcription performed by MT3, this is now possible. To encapsulate this, for each of the 128 MIDI instruments, a one-hot-encoding feature was created with the value of 1 if any note of that instrument is played throughout the audio track and 0 otherwise. Furthermore, the total number of instruments of each specific group present in the song is calculated for each of the five instrument groups, as well as for the binary instrument grouping (melodic vs percussive).
2. **Instrument notes.** This feature represents the number of notes of a specific instrument in a song. The total amount of melodic, percussion, and notes for each of the five instrument groups is also calculated.
3. **Instrument duration percentage.** Based on note duration from MT3 transcription, we calculate the percentage of time an instrument plays throughout each song. For this, the sum of the durations of all notes of a particular instrument is calculated. Then, that amount is divided by the length of the song. Once again, this process was repeated for each of the five instrument groups and the binary groups.

In total, 712 new melody and texture features were created (154 updated features from the additional melodic lines and 558 strictly novel features).

Rhythm Features. As for rhythmic features, our approach relies on the music source separation provided by Demucs [27]. As aforementioned, one of Demucs's

operating modes consists of the separation of a song into the drums stem and the non-drums (melodic) stem. We employed the drums stem as a basis for our novel rhythmic features, as follows.

First, the audio waveform resulting from drums stem is frame-wise analyzed using a 1024-sample frame length, with a hop size of 128 samples (to maintain consistency with the previously extracted features in Panda et al. [22]). Then, several novel features (see the paragraphs below) are extracted from each frame. The sequences resulting from frame-wise feature extraction are then summarized into 6 statistics: mean, standard deviation, skewness, kurtosis, maximum, and minimum. This was performed for the sake of consistency with [22].

In the following, the proposed features are described:

1. **Drum extent percentage.** This feature represents the amount of drums present in the percussion track. First, the Root-Mean-Square (RMS) energy for each frame is computed. Then, the number of frames above a certain threshold is calculated. The ratio between this number of frames and the total amount of frames in the song gives the drum extent of the audio track. This threshold was defined experimentally as 0.025.
   To choose this threshold, a separate small dataset containing 40 songs was employed: 20 excerpts (30-s duration) with low amplitude and percussive content (5 from each quadrant) and 20 excerpts with a high amount of percussion (also 5 from each quadrant) were selected. Then, different threshold values were tested until the calculated drum extent accurately represented the audible amount of drum in the track.
2. **Amplitude and intensity information.** Using the audio waveform, amplitude and intensity information were calculated.
   Amplitude information relies on the full-wave-rectified (FWR) waveform. We start by taking the waveform's absolute value. Then, using the FWR waveform, the six aforementioned statistics were computed.
   On the other hand, intensity information is based on the RMS energy of each frame (as described above). The usual six statistics were also computed.
3. **Spectral features.** A set of spectral features was extracted from the signal in each frame, namely: 1) spectral centroid, 2) spectral bandwidth, 3) spectral contrast, 4) spectral flatness, 5) spectral rolloff, and 6) spectral entropy. The same usual six statistics were computed.
4. **Self-similarity matrix (SSM) features.** Using Mel-Frequency Cepstral Coefficients (MFCC) derived from the audio signal, a set of features that aim to capture information regarding the similarity between each pair of frames in the audio signal is computed. First, the MFCCs are calculated (13 coefficients per frame). Then, the SSM [20] is computed from the MFCC matrix using cosine similarity as the distance metric. Finally, the SSM is thresholded to create a binary matrix, as in (1):

$$\text{Thresholded\_SSM} = \begin{cases} 1 & \text{if SSM} > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here, we denote cells with value one as positive cells corresponding to frames for which a similarity was found. A sequence of positive cells forms a pattern. This SSM serves as a base from which the following features are proposed and extracted:

(a) **Total number of positive cells.** Corresponds to the total number of positive cells where there are similarities.
(b) **Total number of positive cells per second.** The total number of positive cells is divided by the total number of seconds in the song to get the total number of positive cells per second.
(c) **Pattern duration.** An array of pattern durations is computed, where all the repeated patterns' durations are kept. From this array, the six normal statistics are computed.

In total, 57 new rhythmic features were created.

Panda's New Features Extracted on Demucs Stems. Besides the previous features, we also extract Panda's new features from the drums and non-drums stems, after Demucs separation. This amounts to 543 features from each stem (as aforementioned, before feature redundancy analysis, Panda's new feature set comprised 543 features).

Summary of Feature Extraction. Table 2 represents the number of novel features for each proposed feature set.

Table 2. Number of features from each newly proposed feature set.

Feature Set	Num. of Features
Texture and Melody	712
Rhythm	57
Demucs Drums	543
Demucs Non-Drums	543

3.4 Emotion Classification

Due to the high number of features, performing feature reduction is paramount. First, feature redundancy analysis is performed as in [22], and then feature selection is conducted. For this, we employ the ReliefF feature selection algorithm [26] to rank the features based on their importance for emotion classification.

Regarding classification, SVMs [4] were selected as they have been previously employed in our previous work [22].

To tune the hyperparameters, Bayesian search [1] was chosen in detriment of the grid search approach used in our previous work, as it achieved comparable results while taking less time to find the optimal set of parameters. The experiments were then validated using 10 repetitions of 10-fold cross-validation [6], where we report the average (macro-weighted) results (e.g., F1-score).

4 Results and Discussion

In this section, we discuss our classification results. The main objective is to assess the impact of our novel features on the employed dataset. We start by testing the baseline features only, followed by experiments using only the novel features, and, finally, the combination of baseline and novel features. In all comparisons, statistical significance tests are performed ($p < 0.05$).

4.1 Baseline Features Only

Table 3 shows the results on the new datasets using the baseline feature set, where it can be seen that the results are nearly the same in all datasets. However, the MERGE Bimodal Complete dataset required a higher number of features.

Table 3. Results obtained for all datasets using the baseline features only.

Dataset Name	Num. of Features	F1-score
MERGE Audio Complete	200	71.0 ± 2.3
MERGE Audio Balanced	200	70.9 ± 2.3
MERGE Bimodal Complete	300	71.0 ± 2.6
MERGE Bimodal Balanced	200	71.0 ± 2.8

4.2 Novel Features Only

Melody and Texture Features. Using only the texture and melody features, we observe that, while the results are below the ones using baseline features, these F1-scores suggest that these features can aid the MER classification problem. As Table 4 illustrates, we can also start seeing a trend where the bimodal complete dataset has the best results.

Rhythm Features. As Table 5 shows, using only rhythmic features, the results are lower when compared to the other approaches. This can be explained by the fact that percussion is not usually prevalent in some quadrants, e.g., the third quadrant. On the other hand, this aspect suggests the potential of the proposed rhythmic features to discriminate between Q3 and Q4.

Table 4. Results obtained using only the features related to melody and texture.

Dataset Name	# of Features	F1-score
MERGE Audio Complete	300	62.2 ± 2.3
MERGE Audio Balanced	200	62.0 ± 2.6
MERGE Bimodal Complete	200	64.5 ± 2.6
MERGE Bimodal Balanced	200	61.5 ± 3.4

Table 5. Results obtained using the rhythm features.

Dataset Name	Num. of Features	F1-score
MERGE Audio Complete	50	56.6 ± 2.4
MERGE Audio Balanced	40	55.8 ± 2.6
MERGE Bimodal Complete	40	58.7 ± 3.3
MERGE Bimodal Balanced	50	56.6 ± 3.6

All Novel Features Combined. Finally, both sets of features were combined. Table 6 shows the results obtained in this experiment.

As can be observed, the results improve only slightly when compared to the separate approaches. Moreover, in the top 100 features, there are 22 rhythmic features and 78 melodic and texture features. However, the top 5 features are all related to rhythm, with the top feature being Drum Extent Percentage, thus highlighting its importance.

Baseline, Drum and Non-Drum Features. In this experiment, we combine the baseline features from [22] (1255 features) with the Demucs Drums and Demucs Non-Drums features (from Table 2, 543 features each). Table 7 shows the obtained results.

The results obtained do show an increase when compared to the baseline results. This indicates that the isolated melodic part (non-drums stem) of the track can be helpful in MER. This is also supported by the fact that, in the top 150 features, 22 are features extracted from the separated melodic signal. This

Table 6. Results obtained using only the novel features.

Dataset Name	Num. of Features	F1-score
MERGE Audio Complete	300	63.4 ± 2.4
MERGE Audio Balanced	300	63.5 ± 2.4
MERGE Bimodal Complete	200	64.2 ± 3.0
MERGE Bimodal Balanced	200	62.6 ± 3.0

Table 7. Results obtained using the baseline, drum, and non-drum features.

Dataset Name	Num. of Features	F1-score
MERGE Audio Complete	200	72.4 ± 2.2
MERGE Audio Balanced	150	72.4 ± 2.5
MERGE Bimodal Complete	200	74.1 ± 2.5
MERGE Bimodal Balanced	200	72.1 ± 3.0

assumption is further cemented by the fact the difference proved statistically significant for all four datasets ($p < 0.05$).

4.3 Combination of Baseline and Novel Features

Experiments were performed using the novel features combined with the previous baseline features from [22]. Table 8 shows the results obtained for each dataset.

Table 8. Results obtained using the top features with baseline and novel features.

Dataset Name	Num. of Features	F1-score
MERGE Audio Complete	250	72.6 ± 2.4
MERGE Audio Balanced	250	72.8 ± 2.6
MERGE Bimodal Complete	250	74.1 ± 2.5
MERGE Bimodal Balanced	300	72.2 ± 3.0

As can be seen, the overall results show an improvement when adding the novel features to the baseline feature set. The largest increase occurs in the MERGE Bimodal Complete dataset, where the F1-score increased from 71.0% to 74.1%.

It is also important to analyze results obtained per quadrant to observe the main changes. Table 9 and 10 show the confusion matrices for the MERGE Bimodal Complete dataset using only the baseline features and using the combined feature set, respectively.

Table 11 highlights the differences in F1-score from the baseline and the baseline plus novel features for each quadrant.

We can observe an increase in the F1-score in all quadrants, proving the relevance of the newly proposed features for MER studies. The quadrants with the highest increases are the third and the fourth.

In addition, the scores increased when using all the features compared to the baseline results. The major persisting problem is the confusion between Q3 and Q4. However, the novel features helped to improve the discrimination between those quadrants. There is also some confusion between the 1st and 2nd

Table 9. Confusion matrix for the MERGE Bimodal Complete dataset using only baseline features.

	Predicted			
True	**Q1**	**Q2**	**Q3**	**Q4**
	76.32%	10.45%	4.73%	8.50%
	8.70%	89.51%	1.16%	0.64%
	7.86%	2.76%	56.78%	32.60%
	13.15%	0.68%	30.03%	56.15%

Table 10. Confusion matrix for the MERGE Bimodal Complete dataset using baseline plus novel features.

	Predicted			
True	**Q1**	**Q2**	**Q3**	**Q4**
	78.64%	8.31%	4.21%	8.84%
	7.04%	91.47%	1.16%	0.33%
	7.88%	3.32%	60.14%	28.66%
	11.43%	0.52%	27.86%	60.19%

Table 11. Table with the difference in F1-scores between the baseline features and baseline plus novel features for each quadrant.

Dataset Name	Q1	Q2	Q3	Q4
MERGE Audio Complete	+1.3	+0.6	+2.6	+2.4
MERGE Audio Balanced	+1.8	+0.7	+1.9	+3.1
MERGE Bimodal Complete	+2.8	+1.7	+3.3	+4.0
MERGE Bimodal Balanced	+1.5	+0.7	+1.4	+1.3

quadrants, as well as the 1st and 4th, which typically occurs in MER problems. Nevertheless, both of these confusions were reduced by adding the novel features.

Nonetheless, it is also important to analyze the features in the top 200 features for the MERGE Bimodal Complete dataset to further assess the impact of the novel features on the results. Table 12 shows the number of top 200 features that belong to each new feature set. As detailed in previous work [23], these musical elements are emotionally relevant but notably lack audio features.

Table 12. Number of features from each proposed feature set in the top 200 for the MERGE Bimodal Complete dataset.

Feature Set	Num. of Features
Texture and Melody	21
Rhythm	7
Demucs Drums	14
Demucs Non-Drums	20

For the MERGE Bimodal Complete dataset (where the best F1-score of 74.1% was obtained), out of the top-ranked 200 features, 62 were novel. This further confirms the relevance of these features for MER classification. Among these, the new texture features proved particularly relevant, which confirms the hypothesis raised in [23].

5 Conclusions and Future Work

This work proposed several novel features based on the outcomes of music source separation and automatic music transcription frameworks. The newly proposed features helped increase the obtained F1-score, achieving 74.1% with 250 features on the MERGE Bimodal Complete dataset, hence proving their relevance.

The newly proposed features helped mostly to decrease the confusion between the third and fourth quadrants, as well as the confusion between the first and fourth quadrants.

In the future, we propose using deep learning approaches in the MERGE datasets, taking advantage of their larger size compared to other datasets in the literature. In fact, the lack of quality data is one of the biggest drawbacks of these approaches.

Furthermore, we will explore bimodal approaches by combining audio and lyrics, as this is a promising approach to reduce the confusion between Q3 and Q4 since valence information is mostly captured by the lyrics [18].

Finally, additional features could still be developed from the aforementioned tools to help achieve a higher representation of the under-represented musical dimensions.

Acknowledgements. 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. Brownlee, J.: Probability for Machine Learning: Discover How To Harness Uncertainty With Python. Machine Learning Mastery (2019)
2. Cabrera, D., Ferguson, S., Rizwi, F., Schubert, E.: Psysound3: a program for the analysis of sound recordings. J. Acoust. Society America **123**, 3247 (2008). https://doi.org/10.1121/1.2933513
3. Celma, Ò., Herrera, P., Serra, X.: Bridging the Music Semantic Gap, pp. 177–190. Budva, Montenegro (2006)
4. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2** (2007)
5. Dressler, K.: Automatic Transcription of the Melody from Polyphonic Music. Ph.D. thesis, Ilmenau University of Technology (2016)
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley (2000)
7. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. In: 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2003, pp. 375–376. No. 2 in 2, ACM Press, Toronto, Canada (2003). https://doi.org/10.1145/860435.860508

8. Flexer, A., Schnitzer, D., Gasser, M., Widmer, G.: Playlist generation using start and end songs. In: 9th International Society of Music Information Retrieval Conference - ISMIR 2008, pp. 173–178. Philadelphia, Pennsylvania, USA (2008)
9. Friberg, A.: Digital audio emotions - an overview of computer analysis and synthesis of emotional expression in music. In: 11th International Conference on Digital Audio Effects - DAFx 2008, pp. 1–6. Espoo, Finland (2008)
10. Gabrielsson, A.: Emotion perceived and emotion felt: Same or different? Musicae Scientiae **5**(1_suppl), 123–147 (2001). https://doi.org/10.1177/10298649020050S105
11. Gardner, J., Simon, I., Manilow, E., Hawthorne, C., Engel, J.: Mt3: Multi-task multitrack music transcription (2021). https://doi.org/10.48550/ARXIV.2111.03017
12. Hevner, K.: Experimental studies of the elements of expression in music. Am. J. Psychol. **48**(2), 246–268 (1936)
13. Lartillot, O.: MIRtoolbox 1.7.1 User's Manual. University of Oslo (2018)
14. Laurier, C., Herrera, P.: Audio music mood classification using support vector machine. MIREX Task on Audio Mood Classification (2007)
15. Laurier, C., Lartillot, O., Eerola, T., Toiviainen, P.: Exploring relationships between audio features and emotion in music. vol. 3, pp. 260–264. ESCOM, Jyväskylä, Finland (2009)
16. Louro, P.L., Redinho, H., Santos, R., Malheiro, R., Panda, R., Paiva, R.P.: Merge - a bimodal dataset for static music emotion recognition (2024). https://arxiv.org/abs/2407.06060
17. Lu, L., Liu, D.: Automatic mood detection and tracking of music audio signals. Audio, Speech, and Language Processing, IEEE Transactions on **14**, 5–18 (2006). https://doi.org/10.1109/TSA.2005.860344
18. Malheiro, R.: Emotion-Based Analysis and Classification of Music Lyrics. University of Coimbra, Phd (2017)
19. Meyers, O.C.: A Mood-Based Music Classification and Exploration System. Msc, Massachusetts Institute of Technology (2007)
20. Müller, M., Clausen, M.: Transposition-invariant self-similarity matrices. In: 8th International Conference on Music Information Retrieval (ISMIR), pp. 47–50 (2007)
21. Owen, H.: Music theory resource book. Oxford University Press (2000)
22. Panda, R., Malheiro, R., Paiva, R.P.: Novel audio features for music emotion recognition. IEEE Trans. Affect. Comput. **11**(4), 614–626 (2020). https://doi.org/10.1109/TAFFC.2018.2820691
23. Panda, R., Malheiro, R., Paiva, R.P.: Audio features for music emotion recognition: a survey. IEEE Trans. Affect. Comput. **14**(1), 68–88 (2023). https://doi.org/10.1109/TAFFC.2020.3032373
24. Panda, R., Paiva, R.P.: Using support vector machines for automatic mood tracking in audio music. In: 130th Audio Engineering Society Convention 2011 (AES 130), pp. 579–586. Audio Engineering Society, London, UK (2011)
25. Posner, J., Russell, J., Peterson, B.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Develop. Psychopathol. **17**, 715–34 (02 2005). https://doi.org/10.1017/S0954579405050340
26. Robnik-Sikonja, M., Kononenko, I.: An adaptation of relief for attribute estimation in regression. In: International Conference on Machine Learning (1997)
27. Rouard, S., Massa, F., Défossez, A.: Hybrid transformers for music source separation. In: ICASSP 23 (2023)

28. Russell, J.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980). https://doi.org/10.1037/h0077714
29. Sachs, C.: Die hornbostel-sachs'sche klassifikation der musikinstrumente. Naturwissenschaften **2**, 1056–1059 (1914). https://doi.org/10.1007/BF01495319
30. Tzanetakis, G., Cook, P.: Marsyas: a framework for audio analysis. Organised Sound **4** (2002). https://doi.org/10.1017/S1355771800003071
31. Yang, y.h., Lin, Y.C., Su, Y.F., Chen, H.: A regression approach to music emotion recognition. , IEEE Trans. Audio, Speech, Lang. Process. **16**, 448 – 457 (2008). https://doi.org/10.1109/TASL.2007.911513