Novel Audio Features for Music Emotion Recognition

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva

Abstract—This work advances the music emotion recognition state-of-the-art by proposing novel emotionally-relevant audio features. We reviewed the existing audio features implemented in well-known frameworks and their relationships with the eight commonly defined musical concepts. This knowledge helped uncover musical concepts lacking computational extractors, to which we propose algorithms - namely related with musical texture and expressive techniques. To evaluate our work, we created a public dataset of 900 audio clips, with subjective annotations following Russell's emotion quadrants. The existent audio features (baseline) and the proposed features (novel) were tested using 20 repetitions of 10-fold cross-validation. Adding the proposed features improved the F1-score to 76.4 percent (by 9 percent), when compared to a similar number of baseline-only features. Moreover, analysing the features relevance and results uncovered interesting relations, namely the weight of specific features and musical concepts to each emotion quadrant, and warrant promising new directions for future research in the field of music emotion recognition, interactive media, and novel music interfaces.

Index Terms—Affective computing, audio databases, emotion recognition, feature extraction, music information retrieval

1 INTRODUCTION

In recent years, Music Emotion Recognition (MER) has attracted increasing attention from the Music Information Retrieval (MIR) research community. Presently, there is already a significant corpus of research works on different perspectives of MER, e.g., classification of song excerpts [1], [2], emotion variation detection [3], automatic playlist generation [4], exploitation of lyrical information [5] and bimodal approaches [6]. However, several limitations still persist, namely, the lack of a consensual and public dataset and the need to further exploit emotionally-relevant acoustic features. Particularly, we believe that features specifically suited to emotion detection are needed to narrow the so-called semantic gap [7] and their absence hinders the progress of research on MER. Moreover, existing system implementation shows that the state-of-the-art solutions are still unable to accurately solve simple problems, such as classification with few emotion classes (e.g., four to five). This is supported by both existing studies [8], [9] and the small improvements in the results attained in the 2007-2017 MIREX Audio Mood Classification (AMC) task[1], an annual comparison of MER algorithms. These system implementations and research results show a glass ceiling in MER system performances [7].

Several factors contribute to this glass ceiling of MER systems. To begin with, our perception of emotion is inherently subjective: different people may perceive different, even opposite, emotions when listening to the same song. Even when there is an agreement between listeners, there is often ambiguity in the terms used regarding emotion description and classification [10]. It is not well-understood how and why some musical elements elicit specific emotional responses in listeners [10].

Second, creating robust algorithms to accurately capture these music-emotion relations is a complex problem, involving, among others, tasks such as tempo and melody estimation, which still have much room for improvement.

Third, as opposed to other information retrieval problems, there are no public, widely accepted and adequately validated, benchmarks to compare works. Typically, researchers use private datasets (e.g., [11]) or provide only audio features (e.g., [12]). Even though the MIREX AMC task has contributed with one dataset to alleviate this problem, several major issues have been identified in the literature. Namely, the defined taxonomy lacks support from music psychology and some of the clusters show semantic and acoustic overlap [2].

Finally, and most importantly, many of the audio features applied in MER were created for other audio recognition applications and often lack emotional relevance. Hence, our main working hypothesis is that, to further advance the audio MER field, research needs to focus on what we believe is its main, crucial, and current problem: to capture the emotional content conveyed in music through better designed audio features.

This raises the core question we aim to tackle in this paper: which features are important to capture the emotional content in a song? Our efforts to answering this

1. http://www.music-ir.org/mirex/

- *R. Panda and R. P. Paiva are with the Center for Informatics and Systems of the University of Coimbra (CISUC), Coimbra 3004-531, Portugal. E-mail: {panda, ruipedro}@dei.uc.pt.*
- *R. Malheiro is with Center for Informatics and Systems of the University of Coimbra (CISUC) and Miguel Torga Higher Institute, Coimbra 3000-132, Portugal. E-mail: rsmal@dei.uc.pt.*

Manuscript received 10 Jan. 2018; revised 21 Mar. 2018; accepted 24 Mar. 2018. Date of publication 0 . 0000; date of current version 0 . 0000.
(Coressponding author: Renato Panda).
Recommended for acceptance by Y.-H. Yang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2018.2820691

TABLE 1
Musical Features Relevant to MER

Features	Examples
Timing	Tempo, tempo, variation, duration, contrast.
Dynamics	Overall level, crescendo/decrescendo, accents.
Articulation	Overall (staccato, legato), variability.
Timbre	Spectral richness, harmonic richness.
Pitch	High or low.
Interval	Small or large.
Melody	Range (small or large), direction (up or down).
Tonality	Chromatic-atonal, key-oriented.
Rhythm	Regular, irregular, smooth, firm, flowing, rough.
Mode	Major or minor.
Loudness	High or low.
Musical form	Complexity, repetition, disruption.
Vibrato	Extent, range, speed.

question required: i) a review of computational audio features currently implemented and available in the state-of-the-art audio processing frameworks; ii) the implementation and validation of novel audio features (e.g., related with music performance expressive techniques or musical texture).

Additionally, to validate our work, we have constructed a dataset that we believe is better suited to the current situation and problem: it employs four emotional classes, from the Russell's emotion circumplex [13], avoiding both unvalidated and overly complex taxonomies; it is built with a semi-automatic method (AllMusic annotations, along with simpler human validation), to reduce the resources required to build a fully manual dataset.

Our classification experiments showed an improvement of 9 percent in F1-Score when using the top 100 baseline and novel features, while compared to the top 100 baseline features only. Moreover, even when the top 800 baseline features is employed, the result is 4.7 percent below the one obtained with the top100 baseline and novel features set.

This paper is organized as follows. Section 2 reviews the related work. Section 3 presents a review of the musical concepts and related state-of-the-art audio features, as well as the employed methods, from dataset acquisition to the novel audio features and the classification strategies. In Section 4, experimental results are discussed. Finally, conclusions and possible directions for future work are included in Section 5.

2 Related Work

Musical Psychology researchers have been actively studying the relations between music and emotions for decades. In this process, different emotion paradigms (e.g., categorical or dimensional) and related taxonomies (e.g., Hevner, Russell) have been developed [13], [14] and exploited in different computational MER systems, e.g., [1], [2], [3], [4], [5], [6], [10], [11], [15], [16], [17], [18], [19], along with specific MER datasets, e.g., [10], [16], [19].

Emotion in music can be studied as: i) perceived, as in the emotion an individual identifies when listening; ii) felt, regarding the emotional response a user feels when listening, which can be different from the perceived one; iii) or transmitted, representing the emotion that the performer or composer aimed to convey. As mentioned, we focus this work on perceived emotion.

Regarding the relations between emotions and specific musical attributes, several studies uncovered interesting associations. As an example: major modes are frequently related to emotional states such as happiness or solemnity, whereas minor modes are often associated with sadness or anger [20]; simple, consonant, harmonies are usually happy, pleasant or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension or sadness, as they create instability in a musical motion [21]. Moreover, researchers identified many musical features related to emotion, namely: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, rhythm, mode, loudness, vibrato, or musical form [11], [21], [22], [23]. A summary of musical characteristics relevant to emotion is presented in Table 1.

Despite the identification of these relations, many of them are not fully understood, still requiring further musicological and psychological studies, while others are difficult to extract from audio signals. Nevertheless, several computational audio features have been proposed over the years. While the number of existent audio features is high, many were developed to solve other problems (e.g., Mel-frequency cepstral coefficients (MFCCs) for speech recognition) and may not be directly relevant to MER.

Nowadays, most proposed audio features are implemented and available in audio frameworks. In Table 2, we summarize several of the current state-of-the-art (hereafter termed standard) audio features, available in widely adopted frameworks, namely, the MIR Toolbox [24], Marsyas [25] and PsySound3 [26].

Musical attributes are usually organized into four to eight different categories (depending on the author, e.g., [27], [28]), each representing a core concept. Here, we follow an eight categories organization, employing rhythm, dynamics, expressive techniques, melody, harmony, tone colour (related to timbre), musical texture and musical form. Through this organization, we are able to better understand: i) where features related to emotion belong; ii) and which categories may lack computational models to extract musical features relevant to emotion.

One of the conclusions obtained is that the majority of available features are related with tone colour (63.7 percent). Also, many of these features are abstract and very low-level, capturing statistics about the waveform signal or the spectrum. These are not directly related with the higher-level musical concepts described earlier. As an example, MFCCs belong to tone colour but do not give explicit information about the source or material of the sound. Nonetheless, they can implicitly help to distinguish these. This is an example of the mentioned semantic gap, where high level concepts are not being captured explicitly with the existent low level features.

This agrees with the conclusions presented in [8], [9], where, among other things, the influence of the existent audio features to MER was assessed. Results of previous experiments showed that "the used spectral features outperformed those based on rhythm, dynamics, and, to a lesser extent, harmony" [9]. This supports the idea that more adequate audio features related to some musical concepts are lacking. In addition, the number of implemented

TABLE 2
Summary of Standard Audio Features

Name	*RC*	*Description*
Tempo Change	Rhythm	Tempo changes over time.
Beat Spectrum		Measure of acoustic self-similarity.
Onsets		Estimated starting time of the notes.
Events Density		Estimated note onsets per second.
Tempo Estimation		Estimated tempo of the piece.
Fluctuation		Rhythmic periodicity along auditory channels. Estimates rhythm content.
Metrical Analysis		Hierarchical metrical structure info.
Metrical Centroid and Strength		Assessment of metrical activity and pulsation strength / clarity.
Pulse / Rhythm clarity		Strength of the estimated beats.
Beats Loudness		Loudness only for estimated beats.
RMS Energy	Dynamics	The global energy of the signal.
Low Energy Rate		Percentage of frames showing less-than-average energy.
Level		Unweighted sound pressure level of the signal.
Hilbert transform		Instantaneous level, frequency and phase of the audio waveform.
Loudness		Subjective impression of the intensity of a sound.
Timbral Width		The width of the peak of the specific loudness spectrum.
Volume		Refers to the "size" or intensity of the sound.
MaxToTotal		How much the maximum amplitude is off-centre (e.g., crescendos).
TCtoTotal		How the sound is "balanced". (Temporal centroid)/(envelope total length).
Pitch Estimation	Melody	Sequence of continuous pitch values.
Pitch (Terhardt et al.)		Modeling of perceived pitch (outputs several distinct metrics).
Pitch Salience Function		Computes the salience of pitch through time.
Predominant Melody		Estimates F0 of the predominant melody.
Pitch Strength		Indicates if pitch is strongly marked.
Inharmonicity	Harmony	Amount of partials that are not multiples of the F0.
Chromagram		Energy distribution along pitches.
Tuning Frequency		Exact freq. on which a song is tuned.
Key and Key Clarity		Estimated tonal centre positions and their respective clarity.
Key Strength		Probability of each key candidate.
Modality Estimation		Major or minor mode estimation.
Chords Detection and Descriptors		Outputs the sequence of chords in a song and associated descriptors.
Keysom		Chromagram correlation colour map.
Tonal Centroid Vector		6-D tonal centroid from chromagram.
Harmonic Change Detection Function		Flux of the tonal centroid.
Sharpness		Rates sound from dull to sharp.
Spectral & Tonal Dissonance		Harshness among tonal components.

Name	*RC*	*Description*
Attack Time (+ log att)	Tonal Colour	Temporal duration of the attack phases.
Attack Slope & Release		Gradient of attack and release phases.
Attack Leap		Attack phase amplitude.
Avg Notes Duration		Avg duration from attack to release.
Zero Crossing Rate		Waveform sign-change rate.
Spectral Flux		Distance between successive spectral frames.
Spectral Centroid		1st moment (mean): indicates brightness of the sound.
Spectral Spread		2nd moment (variance): measures the dispersion of the spectrum.
Spectral Skewness		3rd moment: symmetry of the spectrum.
Spectral Kurtosis		4th moment: "peakedness" of the data.
Spectral Flatness		Smooth/spikyness of data.
Spectral Contrast		The spectral contrast of a spectrum.
Spectral Entropy		Shannon entropy.
Spectral Rolloff / Brightness		Metrics for the amount of high-frequency energy in the signal.
Bark Bands		Bark band energies (psychoacoustical scale of 24 bands).
Mel Bands		Mel band energies (a perceptual scale of pitches).
ERB Bands		Energies in bands using Equivalent Rectangular Bandwidth scale.
MFCCs		Mel-frequency Cepstral Coefficients – measure of spectral shape.
GFCCs		Gammatone Frequency Cepstral Coeff. – MFCCs using ERB Bands.
LPCC		Linear Predictive Coding Coefficients.
HFC		High-Frequency Content measure.
LSP		Linear Spectral Pairs (coeffs.)
SCF		Spectral Crest Factor, a measure of the "peakiness" of the spectrum.
SFM		Spec. Flatness Measure (inverse of SCF)
Roughness		Estimation of the sensory dissonance (using the peaks of spectrum).
Irregularity		(Successive) spectral peaks variability.
Avg Power Spectrum		Power avg. (over time) of the spectra.
Cepstrum		Inverse Fourier Transform of the log of the spectrum.
Frames Similarity Matrix	MF	Similarity between all possible pairs of frames.
Novelty Curve		Transitions between states.
Average Silence Ratio	ET	Can be used as an assessment of articulation.
Emotion Prediction	Other	Some audio frameworks also provide experimental high-level descriptors based on other lower level features.
Genre Prediction		
Danceability		
Dynamic Complexity		

RC = Related (Musical) Concept; MF =Musical Form; ET = Expressive Techniques

audio features is highly unproportional, with nearly 60 percent in the cited article belonging to timbre (spectral) [9].

In fact, very few features are mainly related with expressive techniques, musical texture (which has none) or musical form. Thus, there is a need for audio features estimating higher-level concepts, e.g., expressive techniques and ornamentations like vibratos, tremolos or staccatos (articulation), texture information such as the number of

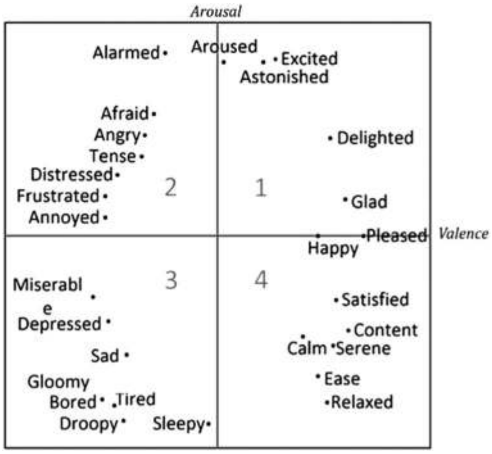

Fig. 1. Russell's circumplex model of emotion (adapted from [9]).

musical lines or repetition and complexity in musical form. Concepts such as rhythm, melody, dynamics and harmony already have some related audio features available. The main question is: are they enough to the problem? In the next sections we address these questions by proposing novel high-level audio features and running classification experiments with both existent and novel features.

To conclude, the majority of current computational MER works (e.g., [3], [10], [16]) share common limitations such as low to average results, especially regarding valence, due to the aforesaid lack of relevant features; lack of uniformity in the selected taxonomies and datasets, which makes it impossible to compare different approaches; and the usage of private datasets, unavailable to other researchers for benchmarking. Additional publicly available datasets exist, most suffering from the same previously described problems, such as: i) Million Song Dataset, which covers a high number of songs but providing only features, metadata and uncontrolled annotations (e.g., based on social media information such as Last. FM) [12]; ii) MoodSwings, which has a limited number of samples [29]; iii) Emotify, which is focused on induced rather than perceived emotions [30]; iv) MIREX, which employs unsupported taxonomies and contains overlaps between clusters [31]; v) DEAM, which is sizeable but shows low agreement between annotators, as well as issues such as noisy clips (e.g., claps, speak, silences) or clear variations in emotion in supposedly static excerpts [32]; vi) or existent datasets, which still require manual verification of the gathered annotations or clips quality, such as [6].

3 Methods

In this section we introduce the proposed novel audio features and describe the emotion classification experiments carried out. To assess this, and given the mentioned limitations of available datasets, we started by building a newer dataset that suits our purposes.

3.1 Dataset Acquisition

The currently available datasets have several issues, as discussed in Section 2. To avoid these pitfalls, the following objectives were pursued to build ours:

1) Use a simple taxonomy, supported by psychological studies. In fact, current MER research is still unable to properly solve simpler problems with high accuracy. Thus, in our opinion, there are few advantages to currently tackle problems with higher granularity, where a high number of emotion categories or continuous values are used;
2) Perform semi-automatic construction, reducing the resources needed to build a sizeable dataset;
3) Obtain a medium-high size dataset, containing hundreds of songs;
4) Create a public dataset prepared to further research works, thus providing emotion quadrants as well as genre, artists or emotion tags for multi-label classification;

Regarding emotion taxonomies, several distinct models have been proposed over the years, divided into two major groups: categorical and dimensional. It is often argued that dimensional paradigms lead to lower ambiguity, since instead of having a discrete set of emotion adjectives, emotions are regarded as a continuum [10]. A widely accepted dimensional model in MER is James Russell's [13] circumplex model. There, Russell affirms that each emotional state sprouts from two independent neurophysiologic systems. The two proposed dimensions are valence (pleasant-unpleasant) and activity or arousal (aroused-not aroused), or AV. The resulting two-dimensional plane forms four different quadrants: 1- exuberance, 2- anxiety, 3- depression and 4- contentment (Fig. 1). Here, we follow this taxonomy.

The AllMusic API[2] served as the source of musical information, providing metadata such as artist, title, genre and emotion information, as well as 30-second audio clips for most songs. The steps for the construction of the dataset are described in the following paragraphs.

Step 1: *AllMusic API querying.* First, we queried the API for the top songs for each of the 289 distinct emotion tags in it. This resulted in 370611 song entries, of which 89 percent had an associated audio sample and 98 percent had genre tags, with 28646 distinct artist tags present. These 289 emotion tags used by AllMusic are not part of any known supported taxonomy, still are said to be "created and assigned to music works by professional editors" [33].

Step 2: *Mapping of AllMusic tags into quadrants.* Next, we use the Warriner's adjectives list [34] to map the 289 AllMusic tags into Russell's AV quadrants. Warriner's list contains 13915 English words with affective ratings in terms of arousal, valence and dominance (AVD). It is an improvement over previous studies (e.g., ANEW adjectives list [35]), with a better documented annotation process and a more comprehensive list of words. Intersecting Warriner and AllMusic tags results in 200 common words, where a higher number have positive valence (Q1: 49, Q2: 35, Q3: 33, Q4: 75).

Step 3: *Processing and filtering*. Then, the set of related metadata, audio clips and emotion tags with AVD values was processed and filtered. As abovementioned, in

2. http://developer.rovicorp.com/docs

our dataset each song is annotated according to one of Russell's quadrants. Hence, the first iteration consisted in removing song entries where a dominant quadrant was not present. We defined a quadrant to be dominant when at least 50 percent of the emotion tags of the song belong to it. This reduced the set to 120733 song entries. Further cleaning was performed by removing duplicated song entries using approximate string matching. A second iteration removed any song entry without genre information and having less than 3 emotion tags associated to meet the predefined objectives, reducing the set to 39983 entries. Then, a third iteration was used to deal with the unbalanced nature of the original data in terms of emotion tags and genres. Finally, the dataset was sub-sampled, resulting in a candidate set containing 2200 song clips, balanced in terms of quadrants and genres in each quadrant, which was then manually validated, as described in the next section.

3.2 Validation of Emotion Annotations

Not many details are known regarding the AllMusic emotion tagging process, apart from supposedly being made by experts [33]. It is unclear whether they are annotating songs using only audio, lyrics or a combination of both. In addition, it is unknown how the 30-second clips that represent each song are selected by AllMusic. In our analysis, we observed several noisy clips (e.g., containing applauses, only speech, long silences, inadequate song segments such as the introduction).

Hence, a manual blind inspection of the candidate set was conducted. Subjects were given sets of randomly distributed clips and asked to annotate them accordingly in terms of Russell's quadrants. Beyond selecting a quadrant, the annotation framework allowed subjects to mark clips as unclear, if the emotion was unclear to the subject, or bad, if the clip contained noise (as defined above).

To construct the final dataset, song entries with clips considered bad or where subjects' and AllMusic's annotations did not match were excluded. The quadrants were also rebalanced to obtain a final set of 900 song entries, with exactly 225 for each quadrant. In our opinion, the dataset dimension is an acceptable compromise between having a bigger dataset using tools such as the Amazon Mechanical Turk or automatic but uncontrolled sources as annotations, and a very small and resource intensive dataset annotated exclusively by a high number of subjects in a controlled environment.

Each song entry is tagged in terms of Russell's quadrants, arousal and valence classes (positive or negative), and multi-label emotion tags. In addition, emotion tags have an associated AV value from Warriner's list, which can be used to place songs in the AV plane, allowing the use of this dataset in regression problems (yet to be demonstrated). Moreover, the remaining metadata (e.g., title, artist, album, year, genre and theme) can also be exploited in other MIR tasks. The final dataset is publicly available in our site[3].

3.3 Standard Audio Features

As abovementioned, frameworks such as the MIR Toolbox, Marsyas and PsySound offer a large number of computational audio features. In this work, we extract a total of 1702 features from those three frameworks. This high amount of features is also because several statistical measures were computed for time series data.

Afterwards, a feature reduction stage was carried to discard redundant features obtained by similar algorithms across the selected audio frameworks. This process consisted in the removal of features with correlation higher than 0.9, where features with lower weight were discarded, according to the ReliefF [36] feature selection algorithm. Moreover, features with zero standard deviation were also removed. As a result, the number of baseline features was reduced to 898. A similar feature reduction process was carried out with the novel features presented in the following subsection.

These standard audio features serve to build baseline models against which new approaches, employing the novel audio features proposed in the next section, can be benchmarked. The illustrated number of novel features is described as follows.

3.4 Novel Audio Features

Many of the standard audio features are low-level, extracted directly from the audio waveform or the spectrum. However, we naturally rely on clues like melodic lines, notes, intervals and scores to assess higher-level musical concepts such as harmony, melody, articulation or texture. The explicit determination of musical notes, frequency and intensity contours are important mechanisms to capture such information and, therefore, we describe this preliminary step before presenting actual features, as follows.

3.4.1 From the Audio Signal to MIDI Notes

Going from audio waveform to music score is still an unsolved problem, and automatic music transcription algorithms are still imperfect [37]. Still, we believe that estimating things such as predominant melody lines, even if imperfect, give us relevant information that is currently unused in MER.

To this end, we built on previous works by Salomon et al. [38] and Dressler [39] to estimate predominant fundamental frequencies (f0) and saliences. Typically, the process starts by identifying which frequencies are present in the signal at each point in time (sinusoid extraction). Here, 46.44 msec (1024 samples) frames with 5.8 msec (128 samples) hopsize (hereafter denoted hop) were selected.

Next, harmonic summation is used to estimate the pitches in these instants and how salient they are (obtaining a pitch salience function). Given this, the series of consecutive pitches which are continuous in frequency are used to form pitch contours. These represent notes or phrases. Finally, a set of computations is used to select the f0s that are part of the predominant melody [38]. The resulting pitch trajectories are then segmented into individual MIDI notes following the work by Paiva et al. [40].

Each of the N obtained notes, hereafter denoted as $note_i$, is characterized by: the respective sequence of f0s (a total of L_i frames), $f0_{j,i},\ j = 1, 2, \ldots L_i$; the corresponding MIDI note numbers (for each f0), $midi_{j,i}$; the overall MIDI note value (for the entire note), $MIDI_i$; the sequence of pitch saliences, $sal_{j,i}$; the note duration, nd_i (sec); starting time, st_i

3. http://mir.dei.uc.pt/resources/MER_audio_taffc_dataset.zip

(sec); and ending time, et_i (sec). This information is exploited to model higher level concepts such as vibrato, glissando, articulations and others, as follows.

In addition to the predominant melody, music is composed of several melodic lines produced by distinct sources. Although less reliable, there are works approaching multiple (also known as polyphonic) F0 contours estimation from these constituent sources. We use Dressler's multi-F0 approach [39] to obtain a framewise sequence of fundamental frequencies estimates.

3.4.2 Melodic Features

Melody is a key concept in music, defined as the horizontal succession of pitches. This set of features consists in metrics obtained from the notes of the melodic trajectory.

MIDI Note Number (MNN) statistics. Based on the MIDI note number of each note, $MIDI_i$ (see Section 3.4.1), we compute 6 statistics: *MIDImean*, i.e., the average MIDI note number of all notes, *MIDIstd* (standard), *MIDIskew* (skewness), *MIDIkurt* (kurtosis), *MIDImax* (maximum) and *MIDImin* (minimum).

Note Space Length (NSL) and *Chroma NSL (CNSL)*. We also extract the total number of unique MIDI note values, *NSL*, used in the entire clip, based on $MIDI_i$. In addition, a similar metric, chroma NSL, *CNSL*, is computed, this time mapping all MIDI note numbers to a single octave (result 1 to 12).

Register Distribution. This class of features indicates how the notes of the predominant melody are distributed across different pitch ranges. Each instrument and voice type has different ranges, which in many cases overlap. In our implementation, 6 classes were selected, based on the vocal categories and ranges for non-classical singers [41]. The resulting metrics are the percentage of MIDI note values in the melody, $MIDI_i$, that are in each of the following registers: Soprano (C4-C6), Mezzo-soprano (A3-A5), Contralto (F3-E5), Tenor (B2-A4), Baritone (G2-F4) and Bass (E2-E4). For instance, for soprano, it comes (1)[4]:

$$RDsoprano = \frac{\sum_{i=1}^{N}[72 \leq MIDI_i \leq 96]}{N}. \tag{1}$$

Register Distribution per Second. In addition to the previous class of features, these are computed as the ratio of the sum of the duration of notes with a specific pitch range (e.g., soprano) to the total duration of all notes. The same 6 pitch range classes are used.

Ratios of Pitch Transitions. Music is usually composed of sequences of notes of different pitches. Each note is followed by either a higher, lower or equal pitch note. These changes are related with the concept of melody contour and movement. They are also important to understand if a melody is conjunct (smooth) or disjunct. To explore this, the extracted MIDI note values are used to build a sequence of transitions to higher, lower and equal notes.

The obtained sequence marking transitions to higher, equal or lower notes is summarized in several metrics, namely: Transitions to Higher Pitch Notes Ratio (*THPNR*), Transitions to Lower Pitch Notes Ratio (*TLPNR*) and Transitions to Equal Pitch Notes Ratio (*TEPNR*). There, the ratio of the number of specific transitions to the total number of transitions is computed. Illustrating for *THPNR*, (2):

$$THPNR = \frac{\sum_{i=1}^{N-1}[MIDI_i < MIDI_{i+1}]}{N-1}. \tag{2}$$

Note Smoothness (NS) statistics. Also related to the characteristics of the melody contour, the note smoothness feature is an indicator of how close consecutive notes are, i.e., how smooth is the melody contour. To this end, the difference between consecutive notes (MIDI values) is computed. The usual 6 statistics are also calculated.

$$NSmean = \frac{\sum_{i=1}^{N-1}|MIDI_{i+1} - MIDI_i|}{N-1}. \tag{3}$$

3.4.3 Dynamics Features

Exploring the pitch salience of each note and how it compares with neighbour notes in the score gives us information about their individual intensity, as well as and intensity variation. To capture this, notes are classified as high (strong), medium and low (smooth) intensity based on the mean and standard deviation of all notes, as in (4):

$$\begin{aligned}
SAL_i &= \underset{1 \leq j \leq L_i}{\text{median}}(sal_{j,i}) \\
\mu_s &= \underset{1 \leq i \leq N}{\text{mean}}(SAL_i) \\
\sigma_s &= \underset{1 \leq i \leq N}{\text{std}}(SAL_i) \\
INT_i &= \begin{cases} low, & SAL_i \leq \mu_s - 0.5\sigma_s \\ medium, & \mu_s - 0.5\sigma_s < SAL_i < \mu_s + 0.5\sigma_s \\ high, & SAL_i \geq \mu_s + 0.5\sigma_s \end{cases}.
\end{aligned} \tag{4}$$

There, SAL_i denotes the median intensity of $note_i$, for all its frames and INT_i stands for the qualitative intensity of the same note. Based on the calculations in (4), the following features are extracted.

Note Intensity (NI) statistics. Based on the median pitch salience of each note, we compute same 6 statistics.

Note Intensity Distribution. This class of features indicates how the notes of the predominant melody are distributed across the three intensity ranges defined above. Here, we define three ratios: Low Intensity Notes Ratio (*LINR*), Medium Intensity Notes Ratio (*MINR*) and High Intensity Notes Ratio (*HINR*). These features indicate the ratio of number of notes with a specific intensity (e.g., low intensity notes, as defined above) to the total number of notes.

Note Intensity Distribution per Second. Low Intensity Note Duration Ratio (LINDR), Medium Intensity Notes Duration Ratio (MINDR) and High Intensity Notes Duration Ratio (HINDR) statistics. These features are computed as the ratio of the sum of the duration of notes with a specific intensity to the total duration of all notes. Furthermore, the usual 6 statistics are calculated.

Ratios of Note Intensity Transitions. Transitions to Higher Intensity Notes Ratio (THINR), Transitions to Lower Intensity Notes Ratio (TLINR) and Transitions to Equal Intensity Notes Ratio (TELNR). In addition to the previous metrics, these features capture information about changes in note

4. Using the Iverson bracket notation.

dynamics by measuring the intensity differences between consecutive notes (e.g., the ratio of transitions from low to high intensity notes).

Crescendo and Decrescendo (CD) statistics. Some instruments (e.g., flute) allow intensity variations in a single note. We identify notes as having crescendo or decrescendo (also known as diminuendo) based on the intensity difference between the first half and the second half of the note. A threshold of 20 percent variation between the median of the two parts was selected after experimental tests. From these, we compute the number of crescendo and decrescendo notes (per note and per sec). In addition, we compute sequences of notes with increasing or decreasing intensity, computing the number of sequences for both cases (per note and per sec) and length crescendo sequences in notes and in seconds, using the 6 previously mentioned statistics.

3.4.4 Rhythmic Features

Music is composed of sequences of notes changing over time, each with a specific duration. Hence, statistics on note durations are obvious metrics to compute. Moreover, to capture the dynamics of these durations and their changes, three possible categories are considered: short, medium and long notes. As before, such ranges are defined according to the mean and standard deviation of the duration of all notes, as in (5). There, ND_i denotes the qualitative duration of $note_i$.

$$\begin{aligned}
\mu_d &= \underset{1 \leq i \leq N}{\text{mean}}(nd_i) \\
\sigma_d &= \underset{1 \leq i \leq N}{\text{std}}(nd_i) \\
ND_i &= \begin{cases} short, & nd_i \leq \ \mu_d - 0.5\sigma_d \\ medium, & \mu_d - 0.5\sigma_d < nd_i < \ \mu_d + 0.5\sigma_d \\ long, & nd_i \geq \ \mu_d + 0.5\sigma_d \end{cases}.
\end{aligned} \tag{5}$$

The following features are then defined.

Note Duration (ND) statistics. Based on the duration of each note, nd_i (see Section 3.4.1), we compute the usual 6 statistics.

Note Duration Distribution. Short Notes Ratio (SNR), Medium Length Notes Ratio (MLNR), Long Notes Ratio (LNR). These features indicate the ratio of the number of notes in each category (e.g., short duration notes) to the total number of notes.

Note Duration Distribution per Second. Short Notes Duration Ratio (SNDR), Medium Length Notes Duration Ratio (MLNDR) and Long Notes Duration Ratio (LNDR) statistics. These features are calculated as the ratio of the sum of duration of the notes in each category to the sum of the duration of all notes. Next, the 6 statistics are calculated for notes in each of the existing categories, i.e., for short notes duration: *SNDRmean* (mean value of *SNDR*), etc.

Ratios of Note Duration Transitions. Ratios of Note Duration Transitions (RNDT). Transitions to Longer Notes Ratio (TLNR), Transitions to Shorter Notes Ratio (TSNR) and Transitions to Equal Length Notes Ratio (TELNR). Besides measuring the duration of notes, a second extractor captures how these durations change at each note transition. Here, we check if the current note increased or decreased in length when compared to the previous. For example, regarding the *TLNR* metric, a note is considered longer than the previous if there is a difference of more than 10 percent in length (with a minimum of 20 msec), as in (6). Similar calculations apply to the *TSNR* and *TELNR* features.

$$TLNR = \frac{\sum_{i=1}^{N-1}[nd_{i+1}/nd_i - 1 > 0.1]}{N-1}. \tag{6}$$

3.4.5 Musical Texture Features

To the best of our knowledge, musical texture is the musical concept with less directly related audio features available (Section 3). However, some studies have demonstrated that it can influence emotion in music either directly or by interacting with other concepts such as tempo and mode [42]. We propose features related with the music layers of a song. Here, we use the sequence of multiple frequency estimates to measure the number of simultaneous layers in each frame of the entire audio signal, as described in Section 3.4.1.

Musical Layers (ML) statistics. As abovementioned, a number of multiple F0s are estimated from each frame of the song clip. Here, we define the number of layers in a frame as the number of obtained multiple F0s in that frame. Then, we compute the 6 usual statistics regarding the distribution of musical layers across frames, i.e., *MLmean*, *MLstd*, etc.

Musical Layers Distribution (MLD). Here, the number of $f0$ estimates in a given frame is divided into four classes: i) no layers; ii) a single layer; iii) two simultaneous layers; iv) and three or more layers. The percentage of frames in each of these four classes is computed, measuring, as an example, the percentage of song identified as having a single layer (*MLD1*). Similarly, we compute *MLD0*, *MLD2* and *MLD3*.

Ratio of Musical Layers Transitions (RMLT). These features capture information about the changes from a specific musical layer sequence to another (e.g., ML1 to ML2). To this end, we use the number of different fundamental frequencies (f0s) in each frame, identifying consecutive frames with distinct values as transitions and normalizing the total value by the length of the audio segment (in secs). Moreover, we also compute the length in seconds of the longest segment for each musical layer.

3.4.6 Expressivity Features

Few of the standard audio features studied are primarily related with expressive techniques in music. However, common characteristics such as vibrato, tremolo and articulation methods are commonly used in music, with some works linking them to emotions [43]–[45].

Articulation Features. Articulation is a technique affecting the transition or continuity between notes or sounds. To compute articulation features, we start by detecting legato (i.e., connected notes played "smoothly") and staccato (i.e., short and detached notes), as described in Algorithm 1. Using this, we classify all the transitions between notes in the song clip and, from them, extract several metrics such as: ratio of staccato, legato and other transitions, longest sequence of each articulation type, etc.

In Algorithm 1, the employed threshold values were set experimentally. Then, we define the following features:

Staccato Ratio (SR), *Legato Ratio (LR)* and *Other Transitions Ratio (OTR).* These features indicate the ratio of each articulation type (e.g., staccato) to the total number of transitions between notes.

Algorithm 1. Articulation Detection.

1. For each pair of consecutive notes, $note_i$ and $note_{i+1}$:
   1.1. Compute the inter-onset interval (*IOI*, in sec), i.e., the interval between the onsets of the two notes, as follows: $IOI = st_{i+1} - st_i$.
   1.2. Compute the inter-note silence (*INS*, in sec), i.e., the duration of the silence segment between the two notes, as follows: $INS = st_{i+1} - et_i$.
   1.3. Calculate the ratio of INS to IOI (*INStoIOI*), which indicates how long the interval between notes is compared to the duration of $note_i$.
   1.4. Define the articulation between $note_i$ and $note_{i+1}$, art_i, as:
      1.4.1. *Legato*, if the distance between notes is less than 10 msec, i.e., $INS \leq 0.01 \Rightarrow art_i = 1$.
      1.4.2. *Staccato*, if the duration of $note_i$ is short (i.e., less than 500 msec) and the silence between the two notes is relatively similar to this duration, i.e., $nd_i < 0.5 \wedge 0.25 \leq INStoIOI \leq 0.75 \Rightarrow art_i = 2$.
      1.4.3. *Other Transitions*, if none of the abovementioned two conditions was met ($art_i = 0$).

Staccato Notes Duration Ratio (SNDR), Legato Notes Duration Ratio (LNDR) and *Other Transition Notes Duration Ratio (OTNDR) statistics.* Based on the notes duration for each articulation type, several statistics are extracted. The first is the ratio of the duration of notes with a specific articulation to the sum of the duration of all notes. Eq. 7 illustrates this procedure for staccato (*SNDR*). Next, the usual 6 statistics are calculated.

$$SNDR = \frac{\sum_{i=1}^{N-1}[\, art_i = 1] \cdot nd_i}{\sum_{i=1}^{N-1} nd_i}. \quad (7)$$

Glissando Features. Glissando is another kind of expressive articulation, which consists in the glide from one note to another. It is used as an ornamentation, to add interest to a piece and thus may be related to specific emotions in music.

We extract several glissando features such as glissando presence, extent, length, direction or slope. In cases where two distinct consecutive notes are connected with a glissando, the segmentation method applied (mentioned in Section 3.4.1) keeps this transition part at the beginning of the second note [40]. The climb or descent, of at least 100 cents, might contain spikes and slight oscillations in frequency estimates, followed by a stable sequence. Given this, we apply the following algorithm:

Then, we define the following features.

Glissando Presence (GP). A song clip contains glissando if any of its notes has glissando, as in (8).

$$GP = \begin{cases} 1, & \text{if } \exists\, i \in \{1, 2, \ldots, N\} : \, gp_i = 1 \\ 0, & \text{otherwise} \end{cases}. \quad (8)$$

Glissando Extent (GE) statistics. Based on the glissando extent of each note, ge_i (see Algorithm 2), we compute the usual 6 statistics for notes containing glissando.

Glissando Duration (GD) and *Glissando Slope (GS) statistics.* As with GE, we also compute the same 6 statistics for glissando duration, based on gd_i and slope, based on gs_i (see Algorithm 2).

Algorithm 2. Glissando Detection.

1. For each note i:
   1.1. Get the list of unique MIDI note numbers, $u_{z,i}$, $z = 1, 2, \ldots, U_i$, from the corresponding sequence of MIDI note numbers (for each f0), $midi_{j,i}$, where z denotes a distinct MIDI note number (from a total of U_i unique MIDI note numbers).
   1.2. If there are at least two unique MIDI note numbers:
      1.2.1. Find the start of the steady-state region, i.e., the index, k, of the first note in the MIDI note numbers sequence, $midi_{j,i}$, with the same value as the overall MIDI note, $MIDI_i$, i.e., $k \;=> \min_{1 \leq j \leq L_i,\; midi_{j,i} = MIDI_i} j$,
      1.2.2. Identify the end of the glissando segment as the first index, e, before the steady-state region, i.e., $e = k - 1$.
   1.3. Define
      1.3.1. gd_i = glissando duration (sec) in note i, i.e., $gd_i = e \cdot hop$.
      1.3.2. gp_i = glissando presence in note i, i.e., $gp_i = 1$ if $gd_i > 0; 0$, otherwise.
      1.3.3. ge_i = glissando extent in note i, i.e., $ge_i = |f0_{1,i} - f0_{e,i}|$ in cents.
      1.3.4. gc_i = glissando coverage of note i, i.e., $gc_i = gd_i / dur_i$.
      1.3.5. $gdir_i$ = glissando direction of note i, i.e., $gdir_i = sign(f0_{e,i} - f0_{1,i})$.
      1.3.6. gs_i = glissando slope of note i, i.e., $gs_i = gdir_i \cdot ge_i / gd_i$.

Glissando Coverage (GC). For glissando coverage, we compute the global coverage, based on gc_i, using (9).

$$GC = \frac{\sum_{i=1}^{N} gc_i \cdot nd_i}{\sum_{i=1}^{N} nd_i}. \quad (9)$$

Glissando Direction (GDIR). This feature indicates the global direction of the glissandos in a song, (10):

$$GDIR = \frac{\sum_{i=1}^{N} gp_i}{N}, when\, gdir_i = 1. \quad (10)$$

Glissando to Non-Glissando Ratio (GNGR). This feature is defined as the ratio of the notes containing glissando to the total number of notes, as in (11):

$$GNGR = \frac{\sum_{i=1}^{N} gp_i}{N}. \quad (11)$$

Vibrato and Tremolo Features. Vibrato is an expressive technique used in vocal and instrumental music that consists in a regular oscillation of pitch. Its main characteristics are the amount of pitch variation (extent) and the velocity (rate) of this pitch variation. It varies according to different music styles and emotional expression [44].

Hence, we extract several vibrato features, such as vibrato presence, rate, coverage and extent. To this end, we apply a vibrato detection algorithm adapted from [46], as follows:

Algorithm 3. Vibrato Detection.

1. For each note i:
   1.1. Compute the STFT, $|\mathrm{F0}_{w,i}|$, $w = 1, 2, \ldots, W_i$, of the sequence $f0_i$, where w denotes an analysis window (from a total of W_i windows). Here, a 371.2 msec (128 samples) Blackman-Harris window was employed, with 185.6 msec (64 samples) hopsize.
   1.2. Look for a prominent peak, $pp_{w,i}$, in each analysis window, in the expected range for vibrato. In this work, we employ the typical range for vibrato in the human voice, i.e., [5], [8] Hz [46]. If a peak is detected, the corresponding window contains vibrato.
   1.3. Define:
      1.3.1. vp_i = vibrato presence in note i, i.e., $vp_i = 1$ if $\exists\ pp_{w,i}$; $vp_i = 0$, otherwise.
      1.3.2. WV_i = number of windows containing vibrato in note i.
      1.3.3. vc_i = vibrato coverage of note i, i.e., $vc_i = WV_i/W_i$ (ratio of windows with vibrato to the total number of windows).
      1.3.4. vd_i = vibrato duration of note i (sec), i.e., $vd_i = vc_i \cdot d_i$.
      1.3.5. $\mathrm{freq}(pp_{w,i})$ = frequency of the prominent peak $pp_{w,i}$ (i.e., vibrato frequency, in Hz).
      1.3.6. vr_i = vibrato rate of note i (in Hz), i.e., $vr_i = \sum_{w=1}^{WV_i} \mathrm{freq}(pp_{w,i})/WV_i$ (average vibrato frequency).
      1.3.7. $|pp_{w,i}|$ = magnitude of the prominent peak $pp_{w,i}$ (in cents).
      1.3.8. ve_i = vibrato extent of note i, i.e., $ve_i = \sum_{w=1}^{WV_i} |pp_{w,i}|/WV_i$ (average amplitude of vibrato).

Then, we define the following features.

Vibrato Presence (VP). A song clip contains vibrato if any of its notes have vibrato, similarly to (8).

Vibrato Rate (VR) statistics. Based on the vibrato rate of each note, vr_i (see Algorithm 3), we compute 6 statistics: *VRmean*, i.e., the weighted mean of the vibrato rate of each note, etc.

$$VRmean = \frac{\sum_{i=1}^{N} vr_i \cdot vc_i \cdot nd_i}{\sum_{i=1}^{N} vc_i \cdot nd_i}. \quad (12)$$

Vibrato Extent (VE) and Vibrato Duration (VD) statistics. As with VR, we also compute the same 6 statistics for vibrato extent, based on ve_i and vibrato duration, based on vd_i (see Algorithm 3).

Vibrato Coverage (VC). Here, we compute the global coverage, based on vc_i, in a similar way to (9).

High-Frequency Vibrato Coverage (HFVC). This feature measures vibrato coverage restricted to notes over note C4 (261.6 Hz). This is the lower limit of the soprano's vocal range [41].

Vibrato to Non-Vibrato Ratio (VNVR). This feature is defined as the ratio of the notes containing vibrato to the total number of notes, similarly to (11).

Vibrato Notes Base Frequency (VNBF) statistics. As with the *VR* features, we compute the same 6 statistics for the base frequency (in cents) of all notes containing vibrato.

As for tremolo, this is a trembling effect, somewhat similar to vibrato but regarding change of amplitude. A similar approach is used to calculate tremolo features. Here, the sequence of pitch saliences of each note is used instead of the f0 sequence, since tremolo represents a variation in intensity or amplitude of the note. Given the lack of scientific supported data regarding tremolo, we used the same range employed in vibrato (i.e., 5-8Hz).

3.4.7 Voice Analysis Toolbox (VAT) Features

Another approach, previously used in other contexts was also tested: a voice analysis toolkit.

Some researchers have studied emotion in speaking and singing voice [47] and even studied the related acoustic features [48]. In fact, "using singing voices alone may be effective for separating the "calm" from the "sad" emotion, but this effectiveness is lost when the voices are mixed with accompanying music" and "source separation can effectively improve the performance" [9].

Hence, besides extracting features from the original audio signal, we also extracted the same features from the signal containing only the separated voice. To this end, we applied the singing voice separation approach proposed by Fan et al. [49] (although separating the singing voice from accompaniment in an audio signal is still an open problem).

Moreover, we used the Voice Analysis Toolkit[5], a "set of Matlab code for carrying out glottal source and voice quality analysis" to extract features directly from the audio signal. The selected features are related with voiced and unvoiced sections and the detection of creaky voice – "a phonation type involving a low frequency and often highly irregular vocal fold vibration, [which] has the potential [...] to indicate emotion" [50].

3.5 Emotion Recognition

Given the high number of features, ReliefF feature selection algorithms [36] were used to select the better suited ones for each classification problem. The output of the ReliefF algorithm is a weight between −1 and 1 for each attribute, with more positive weights indicating more predictive attributes. For robustness, two algorithms were used, averaging the weights: ReliefFequalK, where K nearest instances have equal weight, and ReliefFexpRank, where K nearest instances have weight exponentially decreasing with increasing rank. From this ranking, we use the top N features for classification testing. The best performing N indicates how many features are needed to obtain the best results. To combine baseline and novel features, a preliminary step is run to eliminate novel features that have high correlation with existing baseline features. After this, the resulting feature set (baseline+novel) is used with the same ranking procedure, obtaining a top N set (baseline+novel) that achieves the best classification result.

As for classification, in our experiments we used Support Vector Machines (SVM) [51] to classify music based on the 4 emotion quadrants. Based on our work and in previous MER studies, this technique proved robust and performed generally better than other methods. Regarding kernel selection, a common choice is a Gaussian kernel (RBF),

5. https://github.com/jckane/Voice_Analysis_Toolkit

TABLE 3
Results of the Classification by Quadrants

Classifier	Feat. set	# Features	F1-Score
SVM	**baseline**	70	**67.5% ± 0.05**
SVM	baseline	100	67.4% ± 0.05
SVM	baseline	800	71.7% ± 0.05
SVM	baseline+novel	70	74.7% ± 0.05
SVM	**baseline+novel**	100	**76.4% ± 0.04**
SVM	baseline+novel	800	74.8% ± 0.04

TABLE 5
Confusion Matrix Using the Best Performing Model.

		predicted			
		Q1	*Q2*	*Q3*	*Q4*
actual	*Q1*	185.85	14.40	8.60	18.15
	Q2	23.95	190.55	7.00	3.50
	Q3	14.20	8.40	157.25	45.15
	Q4	24.35	1.65	45.85	153.15
	Total	*246.35*	*215.00*	*218.70*	*219.95*

while a polynomial kernel performs better in a small subset of specific cases. In our preliminary tests RBF performed better and hence was the selected kernel.

All experiments were validated with repeated stratified 10-fold cross validation [52] (using 20 repetitions) and the average obtained performance is reported.

4 Results and Discussion

Several classification experiments were carried out to measure the importance of standard and novel features in MER problems. First, the standard features, ranked with ReliefF, were used to obtain a baseline result. Followingly, the novel features were combined with the baseline and also tested, to assess whether the results are different and statistically significant.

4.1 Classification Results

A summary of the attained classification results is presented in Table 3. The baseline features attained 67.5 percent F1-Score (macro weighted) with SVM and 70 standard features. The same solution achieved a maximum of 71.7 percent with a very high number of features (800). Adding the novel features (i.e., standard + novel features) increased the maximum result of the classifier to 76.4 percent (0.04 standard deviation), while using a considerably lower number of features (100 instead of 800). This difference is statistically significant (at $p < 0.01$, paired T-test).

The best result (76.4 percent) was obtained with 29 novel and 71 baseline features, which demonstrates the relevance of adding novel features to MER, as will be discussed in the next section. In the paragraphs below, we conduct a more comprehensive feature analysis.

Besides showing the overall classification results, we also analyse the results obtained in each individual quadrant (Table 4), which allows us to understand which emotions are more difficult to classify and what is the influence of the standard and novel features in this process. In all our tests, a significantly higher number of songs from Q1 and Q2 were correctly classified when compared to Q3 and Q4. This seems to indicate that emotions with higher arousal are easier to differentiate with the selected features. Out of the two, Q2 obtained the highest F1-Score. This goes in the same direction as the results obtained in [53], and might be explained by the fact that several excerpts from Q2 belong to the heavy-metal genre, which has very distinctive, noise-like, acoustic features.

TABLE 4
Results Per Quadrant Using 100 Features

	baseline			novel		
Quads	*Prec.*	*Recall*	*F1-Score*	*Prec.*	*Recall*	*F1-Score*
Q1	62.6%	73.4%	67.6%	74.6%	81.7%	78.0%
Q2	82.3%	79.6%	80.9%	88.6%	84.7%	86.6%
Q3	61.3%	57.5%	59.3%	71.9%	69.9%	70.9%
Q4	62.8%	57.9%	60.2%	69.6%	68.1%	68.8%

The lower results in Q3 and Q4 (on average 12 percent below the results from Q1 and Q3) can be a consequence of several factors. First, more songs in these quadrants seem more ambiguous, containing unclear or contrasting emotions. During the manual validation process, we observed low agreement (45.3 percent) between the subject's opinions and the original AllMusic annotations. Moreover, subjects reported having more difficulty distinguishing valence for songs with low arousal. In addition, some songs from these quadrants appear to share musical characteristics, which are related to contrasting emotional elements (e.g., a happy accompaniment or melody and a sad voice or lyric). This concurs with the conclusions presented in [54].

For the same number of features (100), the experiment using novel features shows an improvement of 9 percent in F1-Score when compared to the one using only the baseline features. This increment is noticeable in all four quadrants, ranging from 5.7 percent in quadrant 2, where the baseline classifier performance was already high, to a maximum increment of 11.6 percent in quadrant 3, which was the least performing using only baseline features. Overall, the novel features improved the classification generally, with a greater influence in songs from Q3.

Regarding the misclassified songs, analyzing the confusion matrix (see Table 5, averaged for the 20 repetitions of 10-fold cross validation) shows that the classifier is slightly biased towards positive valence, predicting more frequently songs from quadrants 1 and 4 (466.3, especially Q1 with 246.35) than from 2 and 3 (433.7). Moreover, a significant number of songs were wrongly classified between quadrants 3 and 4, which may be related with the ambiguity described previously [54]. Based on this, further MER research needs to tackle valence in low arousal songs, either by using new features to capture musical concepts currently ignored or by combining other sources of information such as lyrics.

4.2 Feature Analysis

Fig. 2 presents the total number of standard and novel audio features extracted, organized by musical concept. As discussed, most are tonal features, for the reasons pointed out previously.

As abovementioned, the best result (76.4 percent, Table 3) was obtained with 29 novel and 71 baseline features, which demonstrates the relevance of the novel features to MER.

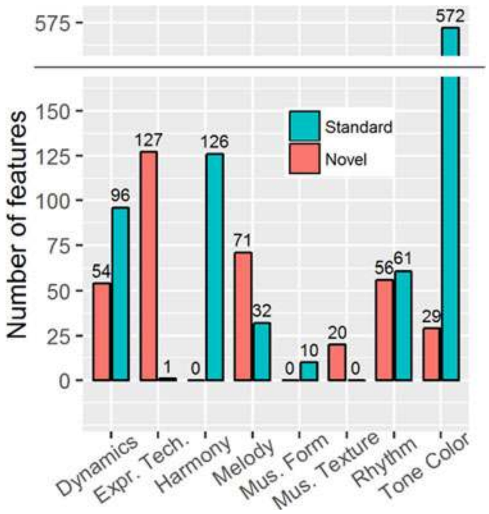

Fig. 2. Feature distribution across musical concepts.

Moreover, the importance of each audio feature was measured using ReliefF. Some of the novel features proposed in this work appear consistently in the top 10 features for each problem and many others are in the first 100, demonstrating their relevance to MER. There are also features that, while alone may have a lower weight, are important to specific problems when combined with others.

In this section we discuss the best features to discriminate each specific quadrant from the others, according to specific feature rankings (e.g., ranking of features to separate Q1 songs from non-Q1 songs). The top 5 features to discriminate each quadrant are presented in Table 6.

Except for quadrant 1, the top5 features for each quadrant contain a majority of tone color features, which are overrepresented in comparison to the remaining. It is also relevant to highlight the higher weight given by ReliefF to the top5 features of both Q2 and Q4. This difference in weights explains why less features are needed to obtain 95 percent of the maximum score for both quadrants, when compared to Q1 and Q3.

Musical texture information, namely the number of musical layers and the transitions between different texture types (two of which were extracted from voice only signals) were also very relevant for quadrant 1, together with several rhythmic features. However, the ReliefF weight of these features to Q1 is lower when compared with the top features of other quadrants. Happy songs are usually energetic, associated with a "catchy" rhythm and high energy. The higher number of rhythmic features used, together with texture and tone color (mostly energy metrics) support this idea. Interestingly, creaky voice detection extracted directly from voice is also highlighted (it ranked 15th), which has previously been associated with emotion [50].

The best features to discriminate Q2 are related with tone color, such as: roughness - capturing the dissonance in the song; rolloff and MFCC – measuring the amount of high frequency and total energy in the signal; and spectral flatness measure – indicating how noise-like the sound is.

Other important features are tonal dissonance (dynamics) and expressive techniques such as vibrato. Empirically, it makes sense that characteristics like sensory dissonance, high energy, and complexity are correlated to tense, aggressive music. Moreover, research supports the association of vibrato and negative energetic emotions such as anger [47].

TABLE 6
Top 5 Features for Each Quadrant Discrimination

Q	Feature	Type	Concept	Weight
Q1	FFT Spectrum - Spectral 2nd Moment (median)	base	Tone Color	0.1467
	Transitions ML1 -> ML0 (Per Sec)	novel	Texture	0.1423
	MFCC1 (mean)	base	Tone Color	0.1368
	Transitions ML0 -> ML1 (Per Sec)	novel (voice)	Texture	0.1344
	Fluctuation (std)	base	Rhythm	0.1320
Q2	FFT Spectrum - Spectral 2nd Moment (median)	base	Tone Color	0.2528
	Roughness (std)	base	Tone Color	0.2219
	Rolloff (mean)	base	Tone Color	0.2119
	MFCC1 (mean)	base	Tone Color	0.2115
	FFT Spectrum - Average Power Spectrum (median)	base	Tone Color	0.2059
Q3	Spectral Skewness (std)	base	Tone Color	0.1775
	FFT Spectrum - Skewness (median)	base	Tone Color	0.1573
	Tremolo Notes in Cents (Mean)	novel	Tremolo	0.1526
	Linear Spectral Pairs 5 (std)	base	Tone Color	0.1517
	MFCC1 (std)	base	Tone Color	0.1513
Q4	FFT Spectrum - Skewness (median)	base	Tone Color	0.1918
	Spectral Skewness (std)	base	Tone Color	0.1893
	Musical Layers (Mean)	novel	Texture	0.1697
	Spectral Entropy (std)	base	Tone Color	0.1645
	Spectral Skewness (max)	base	Tone Color	0.1637

In addition to the tone color features related with the spectrum, the best 20 features for quadrant 3 also include the number of musical layers (texture), spectral dissonance, inharmonicity (harmony), and expressive techniques such as tremolo. Moreover, nine features used to obtain the maximum score are extracted directly from the voice-only signal. Of these, four are related with intensity and loudness variations (crescendos, decrescendos); two with melody (vocal ranges used); and three with expressive techniques such as vibratos and tremolo. Empirically, the characteristics of the singing voice seem to be a key aspect influencing emotion in songs from quadrants 3 and 4, where negative emotions (e.g., sad, depressed) usually have not so smooth voices, with variations in loudness (dynamics), tremolos, vibratos and other techniques that confer a degree of sadness [47] and unpleasantness.

The majority of the employed features were related with tone color, where features capturing vibrato, texture and dynamics and harmony were also relevant, namely spectral metrics, the number of musical layers and its variations, measures of the spectral flatness (noise-like). More features are needed to better discriminate Q3 from Q4, which musically share some common characteristics such as lower tempo, less musical layers and energy, use of glissandos and other expressive techniques.

A visual representation of the best 30 features to distinguish each quadrant, grouped by categories, is represented in Fig. 3. As previously discussed, a higher number of tone

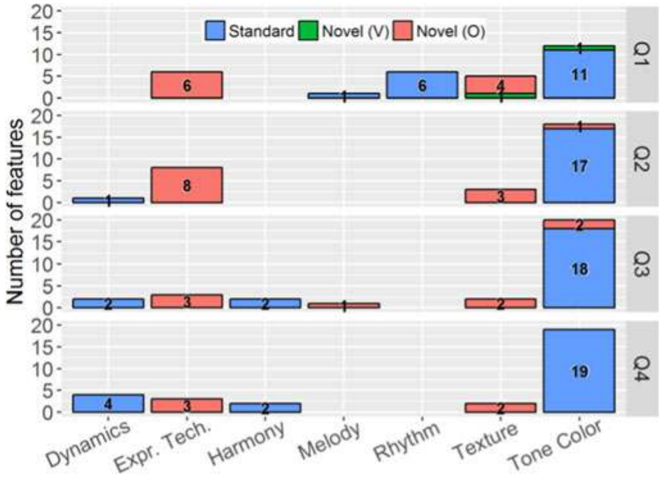

Fig. 3. Best 30 features to discriminate each quadrant, organized by musical concept. Novel (O) are extracted from the original audio signal, while Novel (V) are extracted from the voice-separated signal.

color features is used to distinguish each quadrant (against the remaining). On the other hand, some categories of features are more relevant to specific quadrants, such as rhythm and glissando (part of the expressive techniques) for Q1, or voice characteristics to Q3.

5 Conclusions and Future Work

This paper studied the influence of musical audio features in MER applications. The standard audio features available in known frameworks were studied and organized into eight musical categories. Based on this, we proposed novel more towards higher level musical concepts audio features to help bridge the identified gaps in the state-of-the-art and break the current glass ceiling. Namely, features related with musical expressive performance techniques (e.g., vibrato, tremolo, and glissando) and musical texture, which were the two less represented musical concepts in existing MER implementations. Some additional audio features that may further improve the results, e.g., features related with musical form, are still to be developed.

To evaluate our work, a new dataset was built semi-automatically, containing 900 song entries and respective metadata (e.g., title, artist, genre and mood tags), annotated according to the Russell's emotion model quadrants.

Classification results show that the addition of the novel features improves the results from 67.4 percent to 76.4 percent when using a similar number of features (100), or from 71.7 percent if 800 baseline features are used.

Additional experiments were carried out to uncover the importance of specific features and musical concepts to discriminate specific emotional quadrants. We observed that, in addition to the baseline features, novel features, such as the number of musical layers (musical texture) and expressive techniques metrics, such as tremolo notes or vibrato rates, were relevant. As mentioned, the best result was obtained with 29 novel features and 71 baseline features, which demonstrates the relevance of this work.

In the future, we will further explore the relation between the voice signal and lyrics by experimenting with multi-modal MER approaches. Moreover, we plan to study emotion variation detection and to build sets of interpretable rules providing a more readable characterization of how musical features influence emotion, something that lacks when black-box classification methods such as SVMs are employed.

Acknowledgments

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) – Portugal, as well as the PhD Scholarship SFRH/BD/91523/2012, funded by the Fundação para Ciência e a Tecnologia (FCT), Programa Operacional Potencial Humano (POPH) and Fundo Social Europeu (FSE). The authors would also like to thank the reviewers for their comments that helped improving the manuscript.

References

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, vol. 2, no. 2, pp. 375–376, 2003.

[2] C. Laurier and P. Herrera, "Audio music mood classification using support vector machine," in *Proc. 8th Int. Society Music Inf. Retrieval Conf.*, 2007, pp. 2–4.

[3] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[4] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer, "Playlist generation using start and end songs," in *Proc. 9th Int. Society Music Inf. Retrieval Conf.*, 2008, pp. 173–178.

[5] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-relevant features for classification and regression of music lyrics," *IEEE Trans. Affect. Comput.*, 2016, doi: 10.1109/TAFFC.2016.2598569.

[6] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," in *Proc. 10th Int. Symp. Comput. Music Multidisciplinary Res.*, 2013, pp. 570–582.

[7] Ò. Celma, P. Herrera, and X. Serra, "Bridging the music semantic gap," in *Proc. Workshop Mastering Gap: From Inf. Extraction Semantic Representation*, 2006, vol. 187, no. 2, pp. 177–190.

[8] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Society Music Inf. Retrieval Conf.*, 2010, pp. 255–266.

[9] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimed. Syst.*, pp. 1–25, Aug. 2017, https://link.springer.com/article/10.1007/s00530-017-0559-4

[10] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[11] C. Laurier, "Automatic classification of musical mood by content-based analysis," Universitat Pompeu Fabra, 2011, http://mtg.upf.edu/node/2385

[12] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. 12th Int. Society Music Inf. Retrieval Conf.*, 2011, pp. 591–596.

[13] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[14] K. Hevner, "Experimental studies of the elements of expression in music," *Am. J. Psychol.*, vol. 48, no. 2, pp. 246–268, 1936.

[15] H. Katayose, M. Imai, and S. Inokuchi, "Sentiment extraction in music," in *Proc. 9th Int. Conf. Pattern Recog.*, 1988, pp. 1083–1087.

[16] R. Panda and R. P. Paiva, "Using support vector machines for automatic mood tracking in audio music," in *Proc. 130th Audio Eng. Society Conv.*, vol. 1, 2011, Art. no. 8378.

[17] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Proc. 14th Sound & Music Comput. Conf.*, 2017, pp. 208–213.

[18] N. Thammasan, K. Fukui, and M. Numao, "Multimodal fusion of EEG and musical features music-emotion recognition," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4991–4992.

[19] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS One*, vol. 12, no. 3, Mar. 2017, Art. no. e0173392.

[20] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," in *Music and Emotion*, vol. 8, New York, NY, USA: Oxford University Press, 2001, pp. 223–248.

[21] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen, "Exploring relationships between audio features and emotion in music," in *Proc. 7th Triennial Conf. Eur. Society Cognitive Sciences Music*, vol. 3, 2009, pp. 260–264.

[22] A. Friberg, "Digital audio emotions - An overview of computer analysis and synthesis of emotional expression in music," in *Proc. 11th Int. Conf. Digital Audio Effects*, 2008, pp. 1–6.

[23] O. C. Meyers, *A mood-based music classification and exploration system*. MIT Press, 2007.

[24] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. 10th Int. Conf. Digital Audio Effects (DAFx)*, 2007, pp. 237–244, https://dspace.mit.edu/handle/1721.1/39337

[25] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis," *Organised Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[26] D. Cabrera, S. Ferguson, and E. Schubert, "'Psysound3': Software for acoustical and psychoacoustical analysis of sound recordings," in *Proc. 13th Int. Conf. Auditory Display*, 2007, pp. 356–363.

[27] H. Owen, *Music Theory Resource Book*. London, UK: Oxford University Press, 2000.

[28] L. B. Meyer, *Explaining Music: Essays and Explorations*. Berkeley, CA, USA: University of California Press, 1973.

[29] Y. E. Kim, E. M. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. 9th Int. Society Music Inf. Retrieval Conf.*, 2008, pp. 231–236.

[30] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "Studying emotion induced by music through a crowdsourcing game," *Inf. Process. Manag.*, vol. 52, no. 1, pp. 115–128, Jan. 2016.

[31] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. 9th Int. Society Music Inf. Retrieval Conf.*, 2008, pp. 462–467.

[32] P. Vale, "The role of artist and genre on music emotion recognition," Universidade Nova de Lisboa, 2017.

[33] J. S. Downie, X. Hu, and J. S. Downie, "Exploring mood metadata: Relationships with genre, artist and usage metadata," in *Proc. 8th Int. Society Music Inf. Retrieval Conf.*, 2007, pp. 67–72.

[34] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, Dec. 2013.

[35] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," *Psychology*, vol. Technical, no. C-1, p. 0, 1999.

[36] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003.

[37] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, 2013.

[38] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 6, pp. 1759–1770, Aug. 2012.

[39] K. Dressler, "Automatic transcription of the melody from polyphonic music," Ilmenau University of Technology, 2016.

[40] R. P. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness," *Comput. Music J.*, vol. 30, no. 4, pp. 80–98, Dec. 2006.

[41] A. Peckham, J. Crossen, T. Gebhardt, and D. Shrewsbury, *The Contemporary Singer: Elements of Vocal Technique*. Berklee Press, 2010.

[42] G. D. Webster and C. G. Weir, "Emotional responses to music: interactive effects of mode, texture, and tempo," *Motiv. Emot.*, vol. 29, no. 1, pp. 19–39, Mar. 2005, https://link.springer.com/article/10.1007%2Fs11031-005-4414-0

[43] P. Gomez and B. Danuser, "Relationships between musical structure and psychophysiological measures of emotion," *Emotion*, vol. 7, no. 2, pp. 377–387, May 2007.

[44] C. Dromey, S. O. Holmes, J. A. Hopkin, and K. Tanner, "The effects of emotional expression on vibrato," *J. Voice*, vol. 29, no. 2, pp. 170–181, Mar. 2015.

[45] T. Eerola, A. Friberg, and R. Bresin, "Emotional expression in music: Contribution, linearity, and additivity of primary musical cues," *Front. Psychol.*, vol. 4, 2013, Art. no. 487.

[46] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *IEEE Int. Conf. Acoustics Speech Signal Process.*, 2012, pp. 81–84.

[47] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 218–235, Jan. 2015.

[48] F. Eyben, G. L. Salomão, J. Sundberg, K. R. Scherer, and B. W. Schuller, "Emotion in the singing voice—A deeperlook at acoustic features in the light ofautomatic classification," *EURASIP J. Audio Speech Music Process.*, vol. 2015, no. 1, Dec. 2015, Art. no. 19.

[49] Z.-C. Fan, J.-S. R. Jang, and C.-L. Lu, "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data*, 2016, pp. 178–185.

[50] A. Cullen, J. Kane, T. Drugman, and N. Harte, "Creaky voice and the classification of affect," in *Proc. Workshop Affective Social Speech Signals*, 2013, http://tcts.fpms.ac.be/~drugman/Publi_long/

[51] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[52] R. O. Duda, Peter E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2000.

[53] G. R. Shafron and M. P. Karno, "Heavy metal music and emotional dysphoria among listeners," *Psychol. Pop. Media Cult.*, vol. 2, no. 2, pp. 74–85, 2013.

[54] Y. Hong, C.-J. Chau, and A. Horner, "An analysis of low-arousal piano music ratings to uncover what makes calm and sad music so difficult to distinguish in music emotion recognition," *J. Audio Eng. Soc.*, vol. 65, no. 4, 2017.

Renato Panda received the bachelor's and master's degrees in automatic mood tracking in audio music from the University of Coimbra. He is working toward the PhD degree in the Department of Informatics Engineering, University of Coimbra. He is a member of the Cognitive and Media Systems research group at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests include music emotion recognition, music data mining and music information retrieval (MIR). In October 2012, he was the main author of an algorithm that performed best in the MIREX 2012 Audio Train/Test: Mood Classification task, at ISMIR'2012.

Ricardo Malheiro received the bachelor's and master's degrees (Licenciatura - five years) in informatics engineering and mathematics (branch of computer graphics) from the University of Coimbra. He is working toward the PhD degree at the University of Coimbra. He is a member of the Cognitive and Media Systems research group at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests include natural language processing, detection of emotions in music lyrics and text and text/data mining. He teaches at Miguel Torga Higher Institute, Department of Informatics. Currently, he is teaching decision support systems, artificial intelligence and data warehouses and big data.

Rui Pedro Paiva received the bachelor's, master's (Licenciatura - 5 years) and doctoral degrees in informatics engineering from the University of Coimbra, in 1996, 1999, 2007, respectively. He is a professor with the Department of Informatics Engineering, University of Coimbra. He is a member of the Cognitive and Media Systems research group at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests include music data mining, music information retrieval (MIR) and audio processing for clinical informatics. In 2004, his algorithm for melody detection in polyphonic audio won the ISMIR'2004 Audio Description Contest - melody extraction track, the 1st worldwide contest devoted to MIR methods. In October 2012, his team developed an algorithm that performed best in the MIREX 2012 Audio Train/Test: Mood Classification task.