MUSICAL TEXTURE AND EXPRESSIVITY FEATURES FOR MUSIC EMOTION RECOGNITION

Renato Panda **Ricardo Malheiro** **Rui Pedro Paiva**
CISUC – Centre for Informatics and Systems, University of Coimbra, Portugal
{panda, rsmal, ruipedro}@dei.uc.pt

ABSTRACT

We present a set of novel emotionally-relevant audio features to help improving the classification of emotions in audio music. First, a review of the state-of-the-art regarding emotion and music was conducted, to understand how the various music concepts may influence human emotions. Next, well known audio frameworks were analyzed, assessing how their extractors relate with the studied musical concepts. The intersection of this data showed an unbalanced representation of the eight musical concepts. Namely, most extractors are low-level and related with tone color, while musical form, musical texture and expressive techniques are lacking. Based on this, we developed a set of new algorithms to capture information related with musical texture and expressive techniques, the two most lacking concepts. To validate our work, a public dataset containing 900 30-second clips, annotated in terms of Russell's emotion quadrants was created. The inclusion of our features improved the F1-score obtained using the best 100 features by 8.6% (to 76.0%), using support vector machines and 20 repetitions of 10-fold cross-validation.

1. INTRODUCTION

Music Emotion Recognition (MER) research has increased in the last decades, following the growth of music databases and services. This interest is associated to music's ability to "arouse deep and significant emotions", being "its primary purpose and the ultimate reason why humans engage with it" [1]. Different problems have been tackled, e.g., music classification [2]–[4], emotion tracking [5], [6], playlists generation [7], [8], exploitation of lyrical information and bimodal approaches [9]–[12]. Still, some limitations affect the entire MER field, among which: 1) the lack of public high-quality datasets, as used in other machine learning fields to compare different works; and 2) the insufficient number of emotionally-relevant acoustic features, which we believe are needed to narrow the existing semantic gap [13] and push the MER research forward. Furthermore, both the state-of-the-art research papers (e.g., [14], [15]) and MIREX Audio Mood Classification (AMC) comparison[1] results from 2007 to 2017 are still not accurate enough in easier classification problems with four to five emotion classes, let alone higher granularity solutions and regression approaches, showing a glass ceiling in MER system performances [13].

Many of the audio features applied currently in MER were initially proposed to solve other information retrieval problems (e.g. MFCCs and LPCs in speech recognition [16]) and may lack emotional relevance. Therefore, we hypothesize that, in order to advance the MER field, part of the effort needs to focus on one key problem: the design of novel audio features that better capture emotional content in music, currently left out by existing features.

This raises the core question we aim to tackle in this paper: can higher-level features, namely expressivity and musical texture features, improve emotional content detection in a song?

In addition, we have constructed a dataset to validate our work, which we consider better suited to the current MER state-of-the-art: avoids overly complex or unvalidated taxonomies, by using the four classes or quadrants, derived from the Russell's emotion model [17]; does not require a full manual annotation process, by using AllMusic annotations and data[2], with a simpler human validation, thus reducing resources needed.

We achieved an improvement of up to 7.9% in F1-Score by adding our novel features to the baseline set of state-of-the-art features. Moreover, even when the top 800 baseline features is employed, the result is 4.3% 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 describes the musical concepts and related state-of-the-art audio features. Dataset acquisition, the novel audio features design and classification strategies are also presented. In Section 4, experimental results are discussed. Conclusions and future work are drawn in Section 5.

2. RELATED WORK

Emotions have been a research topic for centuries, leading to the proposal of different emotion paradigms (e.g., categorical or dimensional) and associated taxonomies (e.g.,

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

[2] https://www.allmusic.com/moods

Hevner, Russell) [17], [18]. More recently, these have been employed in many MER computational systems, e.g., [2]–[7], [9], [12], [19], [20], and MER datasets, e.g., [4], [6], [20].

Regarding emotion in music, it can be view as: i) the perceived emotion, identified when listening; ii) emotion felt, representing the emotion felt when listening, which may be different from the perceived; iii) or the emotion transmitted, which is the emotion a performer intended to deliver. This work is focused on perceived emotions, since it is more intersubjective, as opposed to emotion felt, more personal and dependent of context, memories and culture.

As for associations between emotions and musical attributes, many features such as: articulation, dynamics, harmony, loudness, melody, mode, musical form, pitch, rhythm, timbre, timing, tonality or vibrato have been previously linked to emotion [8], [21], [22]. However, many are yet to be fully understood, still requiring further research, while others are hard to extract from audio signals. These musical attributes can be organized into eight different categories, each representing a core concept, namely: dynamics, expressive techniques, harmony, melody, musical form, musical texture, rhythm and tone color (or timbre). Several audio features have been created (hereinafter referred to as standard audio or baseline features) and are nowadays implemented in audio frameworks (e.g. Marsyas [23], MIR toolbox [24] or PsySound [25]). Even though hundreds of features exist, most belong to the same category – tone color, while others were developed to solve previous research problems and thus might not be suited for MER (e.g., Mel-frequency cepstral coefficients (MFCCs) for speech recognition). On the other hand, the remaining categories are underrepresented, with expressivity, musical texture or form nearly absent.

Finally, as opposed to other information retrieval fields, MER researchers lack standard public datasets and benchmarks to compare existent works' adequately. As a consequence, researchers use private datasets (e.g., [26]), or have access only to features and not the actual audio (e.g., [27]). While efforts such as the MIREX AMC task improve the situation, issues have been identified. To begin with, the dataset is private, use in the annual contest only. Also, it uses an unvalidated taxonomy derived from data containing semantic and acoustic overlap [3].

3. METHODS

In this section, due to the abovementioned reasons, we start by introducing the dataset built to validate our work. Following, we detail the proposed novel audio features and emotion classification strategies tested.

3.1 Dataset Creation

To bypass the limitations described in Section 2 we have created a novel dataset based using an accepted and validated psychological model. We decided on Russell's circumplex model [17], which allows us to employ a simple taxonomy of four emotion categories, based on the quadrants resulting from the division by the arousal and valence (AV) axes.

First, we obtained music data (30-second audio clips) and metadata (e.g., artist, title, mood and genre) from the AllMusic API[1]. The mood metadata consisted of several tags per song, from a list of 289 moods. These 289 tags are intersected with the Warriner's list [28] – an improvement on ANEW adjectives list [29], containing 13915 English words with AV ratings according to Russell's model. This intersection results in 200 AllMusic tags mapped to AV, which can be translated to quadrants. Since we considered only songs with three or more mood tags, each song is assigned to the quadrant that has the highest associated number of tags (and at least 50% of the moods are from it).

The AllMusic emotion tagging process is not fully documented, apart from apparently being made by experts [30]. Questions remain on whether these experts are considering only audio, only lyrics or a combination of both. Besides, the 30-second clips selection that represent each song in AllMusic is also undocumented. We observed several inadequate clips (e.g., containing noise such as applauses, only speech, long silences from introductions). Therefore, a manual blind validation of the candidate set was conducted. Subjects were given sets of randomly distributed clips and asked to annotate them according to the perceived emotion in terms of Russell's quadrants.

The final dataset was built by removing the clips where the subjects' and AllMusic derived quadrants' annotations did not match. The dataset was rebalanced to contain exactly 225 clips and metadata per cluster, in a total of 900 song entries, which is publicly available in our site[2].

3.2 Standard or Baseline Audio Features

Marsyas, MIR Toolbox and PsySound3, three state-of-the-art audio frameworks typically used in MER studies, were used to extract a total of 1702 features. This high number is in part due to the computation of several statistical for the resulting time series data. To reduce this and avoid possible feature duplication across different frameworks, first we obtained the weight of each feature to the problem using ReliefF [31] feature selection algorithm. Next, we calculated the correlation between each pair of features, removing the lowest weight one for each pair with a correlation higher than 0.9. This process reduced the standard audio features set to 898 features, which was used to train baseline models. These models were then used to benchmark models trained with the baseline and novel feature sets. An analogous feature reduction procedure was also performed in the novel features set presented in Section 3.3.

3.3 Novel Audio Features

Although being used constantly in MER problems, many of the standard audio features are very low-level, extracting abstract metrics from the spectrum or directly from the audio waveform. Still, humans naturally perceive higher-level musical concepts such as rhythm, harmony, melody

[1] http://developer.rovicorp.com/docs

[2] http://mir.dei.uc.pt/downloads.html

lines or expressive techniques based on clues related with notes, intervals or scores. To propose novel features that related to these higher-level concepts we built on previous works to estimate musical notes and extract frequency and intensity contours. We briefly describe this initial step in the next section.

3.3.1 Estimating MIDI notes

Automatic transcription of music audio signals to scores is still and open research problem [32]. Still, we consider that using such existing algorithms, although imperfect, provide important information currently unused in MER.

To this end, we built on works by Salomon et al. [33] and Dressler [34] to estimate predominant fundamental frequencies (f0) and saliences. This process starts by identifying the frequencies present in the signal at each point in time (sinusoid extraction), using 46.44 msec (1024 samples) frames with 5.8 msec (128 samples) hopsize (hereafter denoted hop). Next, the pitches in each of these moments are estimated using harmonic summation (obtaining a pitch salience function). Then, pitch contours are created from the series of consecutive pitches, representing notes or phrases. Finally, a set of rules is used to select the f0s that are part of the predominant melody [33]. The resulting pitch trajectories are then segmented into individual MIDI notes following the work by Paiva et al. [35].

Each of the N obtained notes, hereafter denoted as $note_i$, is characterized by: 1) the respective sequence of f0s (a total of L_i frames), $f0_{j,i}, j = 1, 2, \dots L_i$; the corresponding MIDI note numbers (for each f0), $midi_{j,i}$; 2) the overall MIDI note value (for the entire note), $MIDI_i$; 3) the sequence of pitch saliences, $sal_{j,i}$; 4) the note duration, nd_i (sec); starting time, st_i (sec); and 5) ending time, et_i (sec). This data is used to model higher level concepts related with expressive techniques, such as vibrato.

In addition to the predominant melody, music typically contains other melodic lines produced by distinct sources. Some researchers have also proposed algorithms to multiple (also known as polyphonic) F0 contours estimation from these constituent sources. We use Dressler's multi-F0 approach [34] to obtain a framewise sequence of fundamental frequencies estimates to assess musical texture.

3.3.2 Musical texture features

Previous studies have verified that musical texture can influence emotion in music, either directly or in combination with tempo and mode [36]. However, as stated in Section 2, very few of the available audio features are directly related with this musical concept. Thus, we propose features to capture information related with the musical layers of a song, based on the simultaneous layers in each frame using the multiple frequency estimates described above.

Musical Layers (ML) statistics. As mentioned, various multiple F0s are estimated from each audio frame. Then, we define the number of layers in a frame as the number of obtained multiple F0s in that frame. The obtained data series, representing the number of musical layers in each instant during the clip, is then summarized using six statistics: mean (MLmean), standard deviation (MLstd), skewness (MLskw), kurtosis (MLkurt), maximum (MLmax) and minimum (MLmin) values. The same six statistics are applied similarly to the other proposed features.

Musical Layers Distribution (MLD). Here, the number of $f0$ estimates in each frame is categorized in one of 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 the song identified as having a single layer (MLD1). Similarly, we compute MLD0, MLD2 and MLD3.

Ratio of Musical Layers Transitions (RMLT). These features capture the amount of transitions (changes) from a specific musical layer sequence to another (e.g., ML1 to ML2). To this end, we count consecutive frames having distinct numbers of fundamental frequencies (f0s) estimated in each as a transition. The total number of these transitions is normalized by the length of the audio segment (in secs). Additionally, we also compute the length in seconds of the longest audio segment for each of the four musical layers classes.

3.3.3 Expressivity features

Expressive techniques such as vibrato, tremolo and articulation are used frequently by composers and performers, across different genres. Some studies have linked them to emotions [37]–[39], still the number of standard audio features studied that are primarily related with expressive techniques is low.

Articulation Features

Articulation relates to how specific notes are played and expressed together. To capture this, we first detect legato (i.e., connected notes played "smoothly") and staccato (i.e., short and detached notes), as defined 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.

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: $IOI = st_{i+1} - st_i$.
   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., $nd_i < 0.5 \land 0.25 \leq INStoIOI \leq 0.75 \Rightarrow art_i = 2$.
      1.4.3. *Other Transitions*, if none of the abovementioned two conditions was met ($art_i = 0$).

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.

Staccato Notes Duration Ratio (SNDR), Legato Notes Duration Ratio (LNDR) and Other Transition Notes Duration Ratio (OTNDR) statistics. These represent statistics based on the duration of notes for each articulation type. As an example, with staccato (SNDR), the ratio of the duration of notes with staccato articulation to the sum of the duration of all notes, as in Eq. 1. For each, the 6 statistics described in Section 3.3.2 are calculated.

$$SNDR = \frac{\sum_{i=1}^{N-1}[art_i = 1] \cdot nd_i}{\sum_{i=1}^{N-1} nd_i} \tag{1}$$

Glissando Features

Glissando is another expressive articulation, which is the slide from one note to another. Normally used as an ornamentation, to add interest to a piece, may be related to specific emotions in music.

We assess glissando by analyzing the transition between two notes, as described in Algorithm 2. This transition part is saved at the beginning of the second note by the segmentation method applied (mentioned in Section 3.3.1) [35]. The second note must start with a climb or descent, of at least 100 cents, which may contain spikes and slight oscillations in frequency estimates, followed by a stable sequence.

ALGORITHM 2
GLISSANDO DETECTION.

1. For each note i:
   1.1. Get the list of unique MIDI note numbers, $u_{z,i}, z = 1, 2, \cdots, U_i$, 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., $k = \min_{1 \le j \le L_i,\, midi_{j,i} = MIDI_i} j$,
      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., $gs_i = gdir_i \cdot ge_i / gd_i$.

Based on the output of Algorithm 2 we define:

Glissando Presence (GP). A song clip contains glissando if any of its notes has glissando, as in (2).

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

If GP = 1, we then compute the remaining glissando features.

Glissando Extent (GE) statistics. Using the glissando extent of each note, ge_i (see Algorithm 2), we compute the 6 statistics (Section 3.3.2) for notes containing glissando.

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

Glissando Coverage (GC). For glissando coverage, we compute the global coverage, based on gc_i, using (3).

$$GC = \frac{\sum_{i=1}^{N} gc_i \cdot nd_i}{\sum_{i=1}^{N} nd_i} \tag{3}$$

Glissando Direction (GDIR). This feature indicates the global direction of the glissandos in a song, (4):

$$GDIR = \frac{\sum_{i=1}^{N} gp_i}{N}, when\ gdir_i = 1 \tag{4}$$

Glissando to Non-Glissando Ratio (GNGR). This feature represents the ratio of the notes containing glissando to the total number of notes, as in (5):

$$GNGR = \frac{\sum_{i=1}^{N} gp_i}{N} \tag{5}$$

Vibrato and Tremolo Features

Vibrato and tremolo are expressive technique used in vocal and instrumental music. Vibrato consists in a steady oscillation of pitch in a note or sequence of notes. Its properties are the: 1) the velocity (rate) of pitch variation; 2) amount of pitch variation (extent); and 3) duration. It varies across music styles and emotional expression [38].

Given its possible relevance to MER, we apply the vibrato detection algorithm described in Algorithm 3, which was adapted from [40]. We then compute features such as vibrato presence, rate, coverage and extent.

ALGORITHM 3
VIBRATO DETECTION.

1. For each note i:
   1.1. Compute the STFT, $|F0_{w,i}|, w = 1, 2, \cdots, W_i$, 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 [40]. 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 (2).

Vibrato Rate (VR) statistics. Based on the vibrato rate value of each note, vr_i (see Algorithm 3), we compute 6 statistics described in Section 3.3.2 (e.g., the vibrato rate weighted mean of all notes with vibrato as in Eq. 6).

$$VRmean = \frac{\sum_{i=1}^{N} vr_i \cdot vc_i \cdot nd_i}{\sum_{i=1}^{N} vc_i \cdot nd_i} \quad (6)$$

Vibrato Extent (VE) and Vibrato Duration (VD) statistics. Similarly to VR, these features represent the same statistics for vibrato extent, based on ve_i and vibrato duration, based on vd_i (see Algorithm 3).

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

Vibrato Coverage (VC). This represents the global vibrato coverage in a song, based on vc_i, similarly to (3).

High-Frequency Vibrato Coverage (HFVC). Here, the VC is computed only for notes over C4 (261.6 Hz), which 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 (5).

An approach similar to vibrato was applied to compute tremolo features. Tremolo can be described as a trembling effect, to a certain degree similar to vibrato but regarding variation of amplitude. Here, instead of using the f0 sequences, the sequence of pitch saliences of each note is used to assess variations in intensity or amplitude. Due to the lack of research regarding tremolo range, we decided to use vibrato range (i.e., 5-8Hz).

3.4 Emotion Classification

Given the high number of features, ReliefF feature selection algorithms [31] were used to rank the better suited ones emotion classification. This algorithm outputs feature weights in the range of -1 to 1, with higher values indicating attributes more suited to the problem. This, in conjunction with the strategy described in Section 3.2, were used to reduce and merge baseline and novel features sets.

For classification we selected Support Vector Machines (SVM) [42] as the machine learning technique, since it has performed well in previous MER studies. SVM parameters were tuned with grid search and a Gaussian kernel (RBF) was selected based on preliminary tests. The experiments were validated with 20 repetitions of 10-fold cross validation [43], where we report the average (macro weighted) results.

4. RESULTS AND DISCUSSION

In this section we discuss the results of our classification tests. Our main objective was to assess the relevance of existing audio features to MER and understand if and how our novel proposed ones improve the current scenario. With this in mind, we start by testing the existing baseline (standard) features only, followed by tests using the combination of baseline and novel, to assess if the obtained results improve and if the differences are statistically significant.

A summary of the classification results is shown in Table 1. The baseline feature set obtained its best result, of 71.7% F1-score, with an extremely high number of features (800). Considering a more reasonable number of features, up to the best 100 according to ReliefF, the best model used the top70, and attained 67.5%. Next, including novel features (with the baseline) increased the best result to 76.0% F1-score using the best 100 features, a considerably lower number (100 instead of 800). This difference is statistically significant (at $p < 0.01$, paired T-test). Interestingly, we observed decreasing results with models using higher number of features, indicating that those extra features might not be relevant but introducing noise.

Classifier	Feature set	# feats.	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	74.0% ± 0.05
SVM	**baseline+novel**	100	**76.0% ± 0.05**
SVM	baseline+novel	800	73.5% ± 0.04

Table 1. Results of the classification by quadrants.

Of the 100 features used in the best result, 29 are novel, which demonstrates the relevance of adding novel features to MER. Of these, 8 are related with texture, such as the number of musical layers (*MLmean*), while the remaining 21 are expressive techniques such as tremolo, glissando and especially vibrato (12). The remaining 71 baseline features are mainly tone color related (50), with the few others capturing dynamics, harmony, rhythm and melody.

Further analysis to the results per individual quadrant, presented in Table 2, gives us a deeper understanding about which emotions are harder to classify and where the new features were more significant. According to it, Q1 and Q2 obtained a higher result compared to the remaining. This seems to indicate that emotions in songs with higher arousal are easier to differentiate. Also, Q2 result is significantly higher, indicating that it might be markedly distinct from the remaining, explained by the fact that several excerpts from Q2 belong to genres such as punk, hardcore or heavy-metal, which have very distinctive, noise-like, acoustic features. This goes in the same direction as the results obtained in previous studies [44].

	baseline			novel		
Quads	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
Q1	62.6%	73.4%	67.6%	72.9%	81.9%	77.2%
Q2	82.3%	79.6%	80.9%	88.9%	82.7%	85.7%
Q3	61.3%	57.5%	59.3%	73.0%	69.2%	71.1%
Q4	62.8%	57.9%	60.2%	68.5%	68.6%	68.5%

Table 2. Results per quadrant using 100 features.

Several factors can be thought to explain the lower results in Q3 and Q4 (average of -11.7%). First, a higher number of ambiguous songs exist in these quadrants, containing unclear or contrasting emotions. This is supported by the low agreement (45.3%) between the subject's and the original AllMusic annotations during the annotation process. In addition, the two quadrants contain songs which share similar musical characteristics, sometimes with each characteristic related to contrasting emotional cues (e.g., a happy melody and a sad voice or lyric). This agrees with the conclusions presented in [45]. As a final point, these similarities may explain why the subjects reported having more difficulty distinguishing valence for songs with low arousal.

The addition of novel features improved the results by 8.6% when considering the top 100 features' results. Novel features seemed more relevant to Q3, with the most significant improvement (by 11.8%), which was before the worst performing quadrant, followed by Q1 (9.6%). On the opposite end, Q2 was already the best performing with baseline features and thus is lower improvement (4.8%).

In addition to assessing the importance of baseline and novel features for quadrants classification, where we identified 29 novel features in the best 100, we also studied the best features to discriminate each specific quadrant from the others. This was done by analyzing specific feature rankings, e.g., the ranking of features that are best to separate Q1 songs from non-Q1 songs (a set containing Q2, Q3 and Q4 annotated as non-Q1). As expected based on former tests, tone color is the most represented concept in the list of the 10 best features for each of the four quadrants. The reason is in part due to being overrepresented in original feature set, while relevant features from other concepts may be missing.

Of the four quadrants, Q2 and Q4 seem to have the most suited features to distinguish them (e.g., features to identify a clip as Q2 vs non-Q2), according to the obtained ReliefF weights. This was confirmed experimentally, where we observed that 10 features or less was enough to obtain 95% of the max score in binary problems for Q2 and Q4, while the top 30 and 20 features, for Q1 and Q3 respectively, were needed to attain the same goal.

Regarding the first quadrant, some of the novel features related with musical texture information were shown to be very relevant. As an example, in the top features, 3 are novel, capturing information related with the number of musical layers and the transitions between different texture types, together with 3 rhythmic features related with events density and fluctuation. Q1 represents happy emotions, which are typically energetic. Associated songs tend to be high in energy and have appealing ("catchy") rhythm. Thus, features related with rhythm, together with texture and tone color (mostly energy metrics) support this. Nevertheless, as stated before the weight of these features to Q1 is low when compared with the top features of other quadrants.

For Q2 the features identified as most suited are related with tone color, such as: roughness - capturing the dissonance in the song; rolloff – measuring the amount of high frequency; MFCCs – total energy in the signal; and spectral flatness measure – indicating how noise-like the sound is. Other important features are related with dynamics, such as tonal dissonance. As for novel features, expressive techniques ones, mainly vibrato, which makes 43% of the top 30 features. Some research supports this association of vibrato and negative energetic emotions such as anger [46]. Generally, the associations found seem reasonable. After all, Q2 is made of tense, aggressive music, and musical characteristics like sensory dissonance, high energy, and complexity are usually present.

Apart from tone color features (extracting energy information), quadrant 3 is also identified higher level features from concepts such as musical texture, dynamics and harmony and expressive techniques. Namely, the number of musical layers, spectral dissonance, inharmonicity, and tremolos. As for quadrant 4, in addition to tone color features related to spectrum (such as skewness or entropy) or measures of how noise-like is the spectrum (spectral flatness), the remaining are again related with dynamics (dissonance) and harmony, as well as some vibrato metrics. More and better features are needed to better understand and discriminate Q3 from Q4. From our tests, songs from both quadrants share some common musical characteristics such as lower tempo, less musical layers and energy, use of glissandos and other expressive techniques.

5. CONCLUSIONS AND FUTURE WORK

We studied the relevance of musical audio features, proposing novel features that complement the existing ones. To this end, the features available in known frameworks were studied and classified in one of eight musical concepts - dynamics, expressive techniques, harmony, melody, musical form, musical texture, rhythm and tone color. Concepts such as musical form, musical texture and expressive techniques were identified as the ones most lacking available audio extractors. Based on this, we proposed novel audio features to mitigate the identified gaps and break the current glass ceiling. Namely, related with expressive techniques, capturing information related with vibrato, tremolo, glissando and articulation. Also, related with musical texture, capturing statistics regarding the musical layers of a musical piece.

Since no public available dataset fulfilled our needs, a new dataset with 900 clips and metadata (e.g., title, artist, genres and moods), annotated according to the Russell's emotion model quadrants was built semi-automatically, used in our tests and is available to other researchers.

Our experimental tests demonstrated that the novel proposed features are relevant and improve MER classification. As an example, using a similar number of features (100), adding our novel proposed features increased the results by 8.6% (to 76.0%), when compared to the baseline. This result was obtained using 29 novel features and 71 baseline, which demonstrates the relevance of this work.

Additional experiments were conducted to uncovered and better understand relations between audio features, musical concepts and specific emotions (quadrants).

In the future, we would like to study multi-modal approaches and the relation between the voice signal and lyrics, as well as testing the features influence in finer grained categorical and dimensional emotion models. Also, other features (e.g. related with musical form), are still to be developed. Moreover, we would like to derive a more understandable set of knowledge (e.g. rules) of how musical features influence emotion, something that lacks when black-box classification methods such as SVMs are employed.

6. ACKNOWLEDGMENT

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 Competitivid-ade (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).

7. REFERENCES

[1] A. Pannese, M.-A. Rappaz, and D. Grandjean, "Metaphor and music emotion: Ancient views and future directions," *Conscious. Cogn.*, vol. 44, pp. 61–71, Aug. 2016.

[2] Y. Feng, Y. Zhuang, and Y. Pan, "Popular Music Retrieval by Detecting Mood," *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, vol. 2, no. 2, pp. 375–376, 2003.

[3] C. Laurier and P. Herrera, "Audio Music Mood Classification Using Support Vector Machine," in *Proc. of the 8th Int. Society for Music Information Retrieval Conf. (ISMIR 2007)*, 2007, pp. 2–4.

[4] 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.

[5] 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.

[6] R. Panda and R. P. Paiva, "Using Support Vector Machines for Automatic Mood Tracking in Audio Music," in *130th Audio Engineering Society Convention*, 2011, vol. 1.

[7] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer, "Playlist Generation Using Start and End Songs," in *Proc. of the 9th Int. Society of Music Information Retrieval Conf. (ISMIR 2008)*, 2008, pp. 173–178.

[8] O. C. Meyers, "A Mood-Based Music Classification and Exploration System," MIT Press, 2007.

[9] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-Relevant Features for Classification and Regression of Music Lyrics," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2016.

[10] X. Hu and J. S. Downie, "When lyrics outperform audio for music mood classification: a feature analysis," in *Proc. of the 11th Int. Society for Music Information Retrieval Conf. (ISMIR 2010)*, 2010, pp. 619–624.

[11] Y. Yang, Y. Lin, H. Cheng, I. Liao, Y. Ho, and H. H. Chen, "Toward multi-modal music emotion classification," in *Pacific-Rim Conference on Multimedia*, 2008, vol. 5353, pp. 70–79.

[12] 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 *10th International Symposium on Computer Music Multidisciplinary Research – CMMR'2013*, 2013, pp. 570–582.

[13] Ò. Celma, P. Herrera, and X. Serra, "Bridging the Music Semantic Gap," in *Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, 2006, vol. 187, no. 2, pp. 177–190.

[14] 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. of the 11th Int. Society for Music Information Retrieval Conf. (ISMIR 2010)*, 2010, pp. 255–266.

[15] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimed. Syst.*, pp. 1–25, Aug. 2017.

[16] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1980.

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

[18] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *Am. J. Psychol.*, vol. 48, no. 2, pp. 246–268, 1936.

[19] 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. of the 14th Sound & Music Computing Conference*, 2017, pp. 208–213.

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

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

[22] A. Friberg, "Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx)*, 2008, pp. 1–6.

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

[24] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction from Audio," in *Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx)*, 2007, pp. 237–244.

[25] D. Cabrera, S. Ferguson, and E. Schubert, "'Psysound3': Software for Acoustical and Psychoacoustical Analysis of Sound Recordings," in *Proc. of the 13th Int. Conf. on Auditory Display (ICAD2007)*, 2007, pp. 356–363.

[26] C. Laurier, "Automatic Classification of Musical Mood by Content-Based Analysis," Universitat Pompeu Fabra, 2011.

[27] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proc. of the 12th Int. Society for Music Information Retrieval Conf. (ISMIR 2011)*, 2011, pp. 591–596.

[28] 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.

[29] 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.

[30] X. Hu and J. S. Downie, "Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata," in *Proc. of the 8th Int. Society for Music Information Retrieval Conf. (ISMIR 2007)*, 2007, pp. 67–72.

[31] 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.

[32] 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.

[33] 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, 2012.

[34] K. Dressler, "Automatic Transcription of the Melody from Polyphonic Music," Ilmenau University of Technology, 2016.

[35] 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.

[36] 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.

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

[38] 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.

[39] T. Eerola, A. Friberg, and R. Bresin, "Emotional expression in music: contribution, linearity, and additivity of primary musical cues.," *Front. Psychol.*, vol. 4, p. 487, 2013.

[40] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 81–84.

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

[42] 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.

[43] R. O. Duda, P. E. (Peter E. Hart, and D. G. Stork, *Pattern classification*. Wiley, 2000.

[44] 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.

[45] 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.

[46] 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.