Audio Features for Music Emotion Recognition: A Survey

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva

Abstract—The design of meaningful audio features is a key need to advance the state-of-the-art in music emotion recognition (MER). This article presents a survey on the existing emotionally-relevant computational audio features, supported by the music psychology literature on the relations between eight musical dimensions (melody, harmony, rhythm, dynamics, tone color, expressivity, texture and form) and specific emotions. Based on this review, current gaps and needs are identified and strategies for future research on feature engineering for MER are proposed, namely ideas for computational audio features that capture elements of musical form, texture and expressivity that should be further researched. Previous MER surveys offered broad reviews, covering topics such as emotion paradigms, approaches for the collection of ground-truth data, types of MER problems and overviewing different MER systems. On the contrary, our approach is to offer a deep and specific review on one key MER problem: the design of emotionally-relevant audio features.

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

1 Introduction

Music Emotion Recognition (MER) is attracting increasing interest from the Music Information Retrieval (MIR) research community. In fact, as pointed out by David Huron nearly 20 years ago, "music's preeminent functions are social and psychological", and so "the most useful retrieval indexes are those that facilitate searching in conformity with such social and psychological functions. Typically, such indexes will focus on stylistic, mood, and similarity information" [1].

There is already a significant corpus of research on different aspects of MER, e.g., classification using symbolic files [2], single-label classification using raw audio excerpts [3], [4], [5], multi-label classification [6], [7], dimensional approaches using regression [8], [9], music emotion variation detection [10], [11], lyrics-based MER [9], bimodal/multi-modal approaches [2], [4], following either classical handcrafted feature design and machine learning [5] or deep learning [10] approaches, with specific MER datasets, e.g., [5], [8], [11]. Nevertheless, several limitations and problems still need to be addressed [5].

- *Renato Panda is with the Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, 3030-290 Coimbra, Portugal, and also with the Ci2, Polytechnic Institute of Tomar, 2300-313 Tomar, Portugal. E-mail: panda@dei.uc.pt.*
- *Ricardo Malheiro is with the Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, 3030-290 Coimbra, Portugal, and also with the Miguel Torga Higher Institute, 3000-132 Coimbra, Portugal. E-mail: rsmal@dei.uc.pt.*
- *Rui Pedro Paiva is with the Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, 3030-290 Coimbra, Portugal. E-mail: ruipedro@dei.uc.pt.*

Manuscript received 11 January 2020; revised 9 October 2020; accepted 13 October 2020. Date of publication 19 October 2020; date of current version 28 February 2023.
(Corresponding author: Rui Pedro Paiva.)
Recommended for acceptance by J. Epps.
Digital Object Identifier no. 10.1109/TAFFC.2020.3032373

Most recent studies have devoted their attention to the MER problems above, datasets and improved machine learning techniques, while applying already existing audio features developed in other contexts, such as speech recognition or music genre classification.

On the other hand, in a previous work [5], we sustained that features specifically suited to emotion detection are needed to narrow the so-called semantic gap [12] and their lack hinders the progress of research on MER. In that work, we designed and implemented novel acoustic features, targeting particularly music expressivity and texture, which led to 9 percent classification improvement (F1-score). Hence, this study supports the argument 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 perspective might as well be transversal to most MIR problems, as pointed out in [13], where the authors affirm that "stagnation on most MIR task results is already acknowledged by MIR community". There, the first hypothesis raised is that "MIR approaches should perhaps be more musical knowledge-intensive" since, currently, mostly generic approaches are followed based on "the application of information retrieval solutions for music, without relying on musically meaningful features" [13]. As Pedro Domingos boldly states, "at the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used" [14].

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 [5], [15] and the small improvements observed in the 2007-2019 Music Information Retrieval Evaluation eXchange (MIREX)[1] Audio Mood Classification

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

task, an annual comparison of MER algorithms. There, the best algorithm achieved 69.8 percent accuracy in a task comprising five categories. Moreover, this score has remained stable for several years, which calls for methods that help breaking the so-called "glass ceiling" [12].

Given the crucial importance of emotionally-relevant audio features for MER, our goal in this survey is threefold:

- to summarize the most significant knowledge on the relations between music and emotion; this review is structured according to eight musical dimensions (melody, harmony, rhythm, dynamics, tone color, expressivity, texture and form) and sets the ground to identify needs in the design of emotionally-relevant audio descriptors;
- to review the current computational audio features that are relevant for MER, particularly the ones available in different open-access audio frameworks, e.g., Marsyas, MIR Toolbox, PsySound and Essentia;
- to unveil possible directions for future research on the topic of feature engineering for MER (based on the above reviews and the identified research needs), as a key effort to break the glass ceiling on audio MER.

Over the years, other authors have offered surveys on Music Emotion Recognition. The most recent we are aware of is the one by Yang *et al.*, from 2018 [15]. Other reviews have been published already several years ago, e.g., the one from 2012 by Yang and Chen [16] or erlier, e.g., [17]. The common characteristic between all of them is that they provide broad MER reviews, tackling topics such as emotion paradigms, approaches for the collection of ground-truth data, types of MER problems (e.g., single-label, multi-label or music emotion variation detection) and overviewing different MER systems. On the contrary, rather than providing a broad but less specific survey, our approach is to offer an updated, deep and specific review on one key MER problem: the design of emotionally-relevant audio features, something that deserved only a somewhat shallow overview in the abovementioned works.

To further clarify the focus of this survey, it is important to mention that approaches based on deep learning techniques are out of the scope of this article, since the breadth of this topic would probably merit a survey in itself. Nevertheless, possible research directions on deep learning for MER are briefly discussed. For the same reason, features based on other modalities, e.g., symbolic or lyrics features, are not covered either. Regarding symbolic features, since some current approaches establish a bridge between the audio and the symbolic MER domains by integrating an audio transcription stage into the feature extraction stage (as discussed in Section 4, e.g., [5]), possible research directions on the exploitation of symbolic features on MER are also briefly discussed.

To summarize, this survey is focused on emotionally-relevant audio features for MER, covering both low-level (e.g., spectral features, MFCC, etc.), perceptual (e.g., rhythm clarity, modality, articulation, etc.) and high-level semantic features (e.g., genre, danceability, etc.) [18], [19].

This paper is organized as follows. Section 2 overviews the relations between music and emotion, which are detailed in Section 3. There, we describe specific associations between each of the eight musical dimensions and different emotions. Section 4 reviews the existing emotionally-relevant computational audio features, organizing them by musical dimension. Section 5 discusses the gaps and needs to advance the study of audio feature design for MER and points directions for future research. Finally, Section 6 concludes the article.

2 Music and Emotion: Overview

Music has been with us since prehistoric times, serving as a language to express our emotions. This is regarded as music's primary purpose [20] and the "ultimate reason why humans engage with it" [21].

Our analysis of the relations between music and emotions is structured according to the fundamental *musical dimensions* usually presented in the musicology literature. Musical dimensions are typically organized into four to eight different categories (depending on the author, e.g., [22], [23]), each representing a core concept. Here, we employ an eight-category organization comprising: melody, harmony, rhythm, dynamics, tone color (or timbre), expressivity, musical texture and musical form.

The organization of these dimensions is not strict. Many musical features are somehow interconnected and may interact and touch other dimensions. Thus, it can be argued that some of them could be placed in different musical categories. In any case, through this organization, we can understand: i) where features related to emotion belong; ii) which features can be extracted from audio signals with the existing algorithms; iii) and thus, which musical dimensions may lack computational models to extract audio features relevant to emotion.

The relations between music and emotions have been debated for millennia, with associations between modes and emotions found in ancient texts, from Indian, Middle Eastern (e.g., Persian), and far eastern (e.g., Japanese) traditions [21]. Natya Shastra (Nāṭya Śāstra), an ancient Sanskrit Hindu text describing performance arts, estimated to have been written somewhere between 500 B.C. and 500 A.D. [24] suggests elements such as modes and musical forms as able to express particular emotions.

In ancient Greece, Plato advocated that "good rhythm wait upon good disposition, [. . .] the truly good and fair disposition of the character and the mind" [25]. In addition, Plato considered harmony as capable of moving the listener, arguing that both "rhythm and harmony find their way to the inmost soul and take strongest hold upon it" [25]. Aristotle supported the same ideas, stating that "rhythms and melodies contain representations of anger and mildness, and also of courage and temperance" [26], while different harmonies could range from relaxing to "violently exciting and emotional" [26].

Scientific studies focusing on the relations between music and emotions started more than a century ago. One of these early examples is a study by Hevner, where the author evaluated the influence of musical factors such as rhythm, pitch, harmony, melody, tempo and mode to each of the eight emotion clusters earlier proposed by her [27]. Along with such studies, music psychologists have proposed different

emotion paradigms (e.g., categorical or dimensional) and related taxonomies (e.g., [27], [28]).

Up to this day, this research problem is still far from completely solved. Nevertheless, several contemporary research works had already identified possible correlations or in some cases causal associations between specific musical elements and emotions. One of the most widely accepted is mode: major modes are frequently related to emotional states such as happiness, whereas minor modes are often associated with sadness or anger [29]; 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 piece [4]. Many other musical elements have been related to emotion, namely, e.g., timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, rhythm, mode, loudness, vibrato or musical form [4], [30].

Over the last decades, several associations have been identified, relating specific emotional responses to the musical dimensions described above. The next section details the most relevant findings in this area. For some musical elements, the research can be somewhat contradicting, which can be caused by many factors, from different research methodologies to differences in the scope of the studies (e.g., induced or perceived emotion, significant differences in methodologies, population, and others). This is also caused by the complexity of the topic and indicates that further research is needed.

Most of the associations that we describe below pertain to music emotion perception[2] or transmission, since most studies tackled that problem. Still, some studies do not clearly state whether their findings concern perceived or induced emotion.

3 Relations Between Musical Dimensions and Emotions

In this section we review the known relations between the eight musical dimensions and different emotions.

3.1 Melody and Emotion

Melody can be defined as a horizontal succession of pitches (perceptual correlate of fundamental frequency), perceived by listeners as a single musical line.

Given its central role in a musical piece, being (one of) the most memorable elements in a song, associations between melodic cues and emotions are expected and suggested since Plato. Some of the strongest relations are found between wider melodic ranges (pitch ranges) and energetic emotions such as joy [31] or fear [32], while narrow ranges are associated with lower arousal emotions, e.g., sadness, melancholy or tranquility [32]. Other melodic elements, such as ascending versus descending melodic contours, have been studied and related to several emotions [27]. However, some of these are disputed in other studies, arguing that the relation is more complex and involves interactions with other elements such as rhythm and modes [32]. These findings have been observed in cross-cultural studies, where listeners have also associated joy with simpler melodies and sadness with more complex ones [31], even when exposed to unfamiliar tonal systems.

TABLE 1
Relations Between Melodic Elements and Emotions

ME	*Value*	*Associated emotions*
Pitch	High	Surprised, angry, fearful, happy [33] and others [32]; increased tense arousal [32]
	Low	Sad, bored, pleasant, increased valence [32]; sad, tender [33]
Pitch variation	Large	Happy, active, surprised [32]; happy [33]
	Small	Angry, bored, disgusted [32]; angry [33]
Pitch range	Wide	Joyful, fearful, scary [32]; happy, fearful [33]
	Narrow	Sad [32]; sad, tender [33]
Melodic intervals	Large	Powerful [34]
	Minor 2nd	Melancholic [34], sad [33]
	Perfect 4th, major 6th, minor 7th	Carefree [32]; happy (perfect 4th) [33]
	Perfect 5th	Carefree, active [32], happy [33]
	Octave	Carefree, positive, strong [32]
Melodic direction and contour	Ascending	Happy, fearful, surprised, angry, tense [32]
	Descending	Sad, bored, pleasant [32]
Melodic movement	Stepwise motion	Dull melodies [35]
	Intervallic leaps or skips	Exciting melodies [35]
	Stepwise and skipwise leaps	Peaceful melodies [35]

Table 1 summarizes the known relations between melody and emotion. There, ME stands for Musical Element.

3.2 Harmony and Emotion

If melody is said to be the horizontal part of music, harmony refers to its vertical aspect, i.e., the sound produced by the combination of various pitches (notes or tones) in chords.

Harmony, together with rhythm and melody, was thought as able to elicit emotions since ancient times. Consonant harmonies are usually associated with happiness, tranquility, serenity, while dissonant complex harmonies are related with negative emotional states, e.g., tension and sadness, due to the instability they create in the piece [4].

In addition, major modes have been frequently related with positive emotions (e.g., happiness), while minor modes are linked to negative ones (e.g., sadness) [32]. Some authors such as Cook *et al.* have tried to further understand this affective response to major/minor chords and resolved/unresolved chords, concluding that this emotional association is "neither due to the summation of interval effects nor simply arbitrary, learned cultural artifacts, but rather that harmony has a psychophysical basis dependent on three-tone combinations" [36].

2. Emotion in music can be regarded as: i) perceived, as in the emotion an individual identifies when listening; ii) induced or 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 [8].

TABLE 2
Relations Between Harmony and Emotions

ME	Value	Associated emotions
Harmonic perception (harmonic intervals)	Consonant (simple)	Normally associated with positive emotions, e.g., happy [33], serene and dignified [27], pleasant, tender [32]
	Dissonant (complex)	Associated mostly with negative emotions: vigorous, sad [27][33], unpleasant, tense, fearful, angry [32]
	High-pitched	Happy, more active/powerful [32]
	Low-pitched	Sad, less powerful [32]
Harmony (tonality)	Tonal	In joyful, dull or peaceful melodies, pleasant [32]
	Atonal	In angry melodies [32][33]
	Using chromatic scales	In sad and angry melodies [32]
Harmony (mode)	Major	Positive emotions, e.g., happy, serene, tender [32]; happy [33]
	Minor	Negative emotions, e.g., sad, disgusted and angry [32]; sad [33]

The relations between harmony and emotion are summarized in Table 2.

3.3 Rhythm and Emotion

Rhythm represents the element of "time" in music, the patterns of long and short sounds and silences found in music.

Rhythm, together with melody and harmony, is one of the dimensions most associated with the emotional expression in music. In fact, some authors consider it the most important one, e.g., [37], [38]. Rhythm elements, such as the augmentation of tempo (from 90 to 150 bpm), has been shown to increase happiness and surprise measures (i.e., induce) [39], while decreasing sadness. In the study, the authors used two groups of words to study different emotion types: 3 "basic emotions" where users reported what they felt (i.e., induced emotion) on a scale of 1 to 8; and 4 "descriptive words" (tension, expressiveness, amusement and attractiveness) to classify (i.e., perceived emotion) the musical piece on a scale of 1 to 5.

In addition to tempo, the rhythmic unit of a piece has also been shown to influence the emotional message of a song. As an example, variations "of the rhythm of the melody without altering the musical line, harmonics or beat" [39], such as changes from whole and half notes (theme) to eighth or sixteenth, as well syncopated notes, were associated with specific emotions. Similar studies have supported the idea that rhythm is somehow influencing the emotional information in music, e.g., [40].

The associations between rhythm and emotion are summarized in Table 3, based on the reviews presented in [32], [33], [41], as well as the other mentioned papers.

TABLE 3
Relations Between Rhythm and Emotion

ME	Value	Associated emotions
Tempo	Fast	Several, among which happy, graceful, vigorous, pleasant, active, angry, fearful, energy arousal and tension arousal [32]; happy, anger, fear [33]; high arousal, e.g., happy, stressful, amusing [39]
	Slow	Several, among which serene, dreamy, dignified, serious, tranquil, sentimental, dignified, sad, peaceful [32]; sad, tender [33]
Tempo and Note Values	High tempo (150 bpm) and sixteenth notes	High arousal: happy, amusing, expressive, stressful [39]
	Moderate to fast tempo (120 or 150 bpm) and sixteenth notes	Surprised [39]
	Slow to moderate tempo (90 bpm) and whole and half notes	Sad, boring, relaxing, expressionless [39]
Rhythm Types	Regular/smooth	Happy, glad, serious, dignified, peaceful, majestic [32]; happy, anger [33]
	Irregular/rough	Amusing, uneasy [32]
	Complex	Angry [32] [33]
	Varied	Joyful [32]; fear [33]
	Firm	Dignified, vigorous, sad, exciting[3] [27], sad [32]
	Flowing/fluent	Happy, dreamy, graceful, serene [27], gay [32]
Rests	After tonal closure (a sequence which starts and ends in the same key)	Lower tension [32]
	After no tonal closure	Higher tension than observed if after tonal closure [32]

3.4 Dynamics and Emotion

Dynamics represents the variation in loudness or softness of notes in a musical piece.

The influence of dynamics, namely loudness and loudness variations, in music emotions (both induced and perceived) have been studied by some researchers, some of which relate them with specific emotion states. Empirically, an association of loud music (high intensity) with powerful and intense emotions such as joy, anger or tension seems logical. In contrast, soft music is mostly linked to calm, serene or sad music. Such associations have been verified by several researchers [42], [38], [43]. Variations in loudness over a musical piece have also been studied. Namely, larger variations are usually more negative [43], while smaller variations are more positive [32].

Table 4 summarizes the associations between dynamics and emotion.

3.5 Tone Color and Emotion

Tone color (or timbre) is related to lower level elements and properties of the sound itself, e.g., amplitude and spectrum, essential to differentiate instruments and voices.

Several sound properties have been associated with emotional states. A rounder amplitude envelope is related with

TABLE 4
Relations Between Dynamics and Emotion

ME	Value	Associated emotions
Dynamic levels (forte, piano, etc.)	High/Loud	Excited, triumphant, strong/powerful, tense, angry, energy arousal and tension arousal [32]; happy, anger [33]
	Low/Soft	Melancholic, peaceful, solemn, fearful, tender, sad, lower intensity, higher valence [32]; sad, fear, tender [33]
Accents and changes in dynamic levels	Large	Fearful [32] [33]
	Small	Happy [33], pleasing, active [32]
	Rapid variations	Playful, pleading, fearful [32]
	No changes	Sad, peaceful, dignified, happy [32]
	Crescendo, decrescendo, accelerando, ritardando	Said to be useful to describe perceptual and emotional processes [44]; anger (accelerando) [33]

TABLE 5
Relations Between Tone Color (Timbre) and Emotion

ME	Value	Associated emotions
Amplitude envelope	Round	Disgusted, bored, potent, fear, sadness [32]
	Sharp	Pleasant, happy, surprised, active, angry [32]; angry [33]
Spectral envelope (no. harmonics)	Low	Bored, happy, pleasant, sad [32]
	High	Active, angry, disgusted, fearful, potent, surprised [32]
Spectral characteristics (e.g., spectral centroid, etc.)	Positive correlation	Positive emotions: happy, heroic, comic, joyful [45, 47]
	Negative correlation	Negative emotions: sad, scary, shy, depressed [45, 47]

negative emotions such as disgust, sadness or fear [32], [38], while a sharper one gives rise to positive emotions such as happiness or surprise [32], with some authors also linking it to fear [38]. The number of harmonics has also been studied, where a lower number is associated with boredom, happiness or sadness [32], while a high number of harmonics is usually related with emotions with high arousal and negative valence, e.g., anger, disgust, fear [32].

The tone color of specific instruments has also been suspected to carry emotional expression cues. In fact, composers and movie and marketing directors select specific instruments to express distinct emotions. This idea has been supported by studies such as [45]. In this respect, Hailstone *et al.* state that "timbre (instrument identity) independently affects the perception of emotions in music after controlling for other acoustic, cognitive, and performance factors" [46]. These works highlight the importance of spectral centroid (brightness) as a "significant component in music emotion". Moreover, spectral centroid deviation, spectral shape, attack time and even/odd harmonic ratio were all considered relevant [45].

A summary of the relations between tone color and emotion is presented in Table 5.

3.6 Expressivity and Emotion

Expressive techniques in music encompass several ornaments and features that are used by both composers (to enrich their pieces) and performers (to express their emotions at specific moments). Both parts have been studied and related with specific emotional states. As an example, staccato articulation is normally associated with higher intensity and energetic [32], mostly negative as with fear and anger [38]. On the other hand, legato is associated with softness [32] and sadness [38]. Similar research has been conducted regarding vibratos and emotion expression, observing that "singing an emotional passage influences acoustic features of vibrato when compared with isolated, sustained vowels" [48]. To assess this, classical singers were asked to sing passages of their preference containing both high and low levels of emotion. The analysis of the recordings shows significant changes in vibrato characteristics such as frequency modulation rate and extent.

Regarding emotion expression by the performer, some studies highlighted that artists typically use different ornaments, such as accentuating specific notes considered happy, whereas not doing the same for sadness [49]. In addition, Timmers and Ashley studied the usage by flute and violin performers of specific ornamentations such as trills, turns, mordente, arpeggio and others, when they intended to express one of four specific affect terms (happiness, sadness, anger and love), and how these emotions were perceived by listeners [50]. The accuracy between intended versus rated emotions was lowest for happiness. The performers employed more complex ornamentations for angry and the least complex for sadness.

Table 6 summarizes the main relations between expressivity and emotion.

3.7 Texture and Emotion

Musical texture refers to the way the rhythmic, melodic and harmonic information produced by musical instruments and voices is combined in a musical composition. It is thus related to the combination and relations between the musical lines or layers (one or more instruments with the same role) in a song.

Fewer studies have been conducted regarding musical texture and emotions and of these some contain contradicting results. In one of the oldest studies, Kastner and Crowder evaluated the emotional differences between monophonic (melody only) and homophonic textures (melody with block chords accompaniment) by children aged three to twelve. In that study, the unaccompanied version (monophonic) was rated as more positive [51]. A similar result was observed by Webster and Weir, where nonharmonized melodies were considered happier [52]. However, further studies attempting to replicate Kastner and Crowder's findings observed exactly the opposite result. There, not only children but also adult subjects considered monophonic sounds as less happy than accompanied ones [53], [54]. A possible explanation to this contradicting results are the different versions of "dense textures" used in each [55], where very basic/simple chords and a single instrument were used in the studies observing negative emotions, while the others used more complex (and thus, with higher density) accompaniments taken from published songbooks. These differences may influence greatly

TABLE 6
Relations Between Expressivity and Emotion

ME	Value	Associated emotions
Articulation	Legato	Soft [32], tender, sad [32][33]
	Staccato	Intense, energetic, active, fearful, angry [32]; happy [33]
Ornamentation[4]	Single appoggiatura	[pos.] Flute: lovely, sad [50] [neg.] Flute: happy, angry [50]
	Double appoggiatura	[neg.] Violin: sad [50]
	Trill	[pos.] Flute: angry [50] [neg.] Flute: lovely, sad [50]
	Turn	[pos.] Violin: happy [50]
	Mordent	No significant correlation was observed [50]
	Slide	No significant correlation was observed [50]
	Arpeggio	[pos.] Flute: angry [50] [neg.] Flute: lovely, sad [50]
	Substitute	[pos.] Violin: sad [50]
Vibrato	Higher frequency modulation (FM) rate + higher FM extent + lower modulation variability	Observed when classical singers sang "more emotional passages"[5] (as opposed to neutral songs) [48]; Happy (medium-fast rate, medium extent) [33]
	Higher mean F_0 + higher mean intensity	Observed in "more emotional passages" [48]

other musical dimensions (e.g., harmony) making it harder to correctly compare the results.

Polyphonic textures, containing several voices, have also been explored recently, suggesting that music with a higher number of voices is perceived as more positive. Such musical excerpts were rated as "sounding more happy, less sad, less lonely, and more proud" [55].

Although further studies are required to better understand exactly how musical texture influences emotion, the existing ones have demonstrated that it can indeed influence emotion in music either directly or by interacting with other features such as tempo and mode [55].

Table 7 summarizes the associations between musical texture and emotions.

3.8 Form and Emotion

Musical form or musical structure refers to the overall structure of a musical piece and describes the layout of a composition as divided into sections.

Some studies have investigated possible relations between musical form and emotion. It seems that forms with lower complexity are associated with positive emotions [56] such as relaxation, joy or peace [31]. On the contrary, higher complexity forms usually result in more negative emotions such as sadness [31], which can be higher in arousal (e.g., aggressive) or lower (e.g., melancholy) depending on the dynamism (high or low, respectively) [56].

Some researchers explored the relation between emotion and form by changing the order of sections (in classical music) but no relevant results were obtained [57], [58].

TABLE 7
Relations Between Texture and Emotion

ME	Value	Associated emotions
Texture type	Monophonic	More positive [51] and happier [52] than homophonic
	Homophonic	Happier [53, 54] than monophonic.
Number of layers and density	Music with higher number of voices (polyphonic)	"more happy, less sad, less lonely, and more proud" [55]

The few associations found between musical form and emotions are presented in Table 8.

3.9 Interactions Between Musical Dimensions

As described in the previous sections, each musical element may influence distinct emotional expressions. In fact, emotional content in music is not defined exclusively by a single element but is built by the merging and interaction of several factors. Beyond studying associations concerning musical dimensions and emotions independently, these interactions between several musical dimensions and the associated emotional responses have also been studied and reviewed, e.g., [59], [60].

Such works unveil interesting indirect relations and interactions regarding the variation of specific elements and the corresponding emotional changes, as well as possible interactions between elements, resulting in different emotional states. One example is the interaction between tempo and mode [60]: high tempo and minor mode results in only high arousal, while the same high tempo, but with major mode, results in high arousal and positive valence.

Several other authors have studied possible interactions, such as mode and tempo [37], the influence of pitch height, intensity and tempo in valence [42], the influence of rhythm, melodic contour and melodic progression in happy music [32] or interactions between tempo, texture and mode [52].

4 Computational Audio Features in MER

In general terms, a feature is a characteristic part of something. Features help to distinguish one thing from another, by providing the essential descriptive primitives by which individual objects or works may be identified [61].

In musical terms, features may be characteristic of a musical work, of a movement, of a composer, of a very specific musical dimension, of a genre, and so forth. As Huron states, "what constitutes a feature depends on the scope of

TABLE 8
Relations Between Form and Emotion

ME	Value	Associated emotions
Form complexity	Low	Positive emotions [56], Joy, peace, relaxation [31]
	High	Sadness [31]
	High complexity and low dynamism	Depression, melancholy [56]
	High complexity and high dynamism	Aggressiveness, anxiety [56]

our gaze" [61]. For illustration, features can be employed to represent any aspect that is relevant to the identification of a song, from the chords, to abstract statistics regarding physical aspects of the sound wave, rhythm information and others. Summing it up, the goal of feature extraction is to reduce the information of songs to descriptors that can accurately describe them [15].

Over the last decades, several algorithms have been proposed to extract information from audio signals. These features have been developed to solve a myriad of problems, from speech recognition, to content-based retrieval, indexing, and fingerprinting. More recently, a few works studied how the human perception of music characteristics (e.g., tempo) correlates with these audio descriptors, e.g., [62], [63]. It was observed that some features, "in particular those related to loudness, timbre, harmony, and rhythm show high correlations with perceived emotions" [63]. Still, such studies are usually carried with small datasets or specific genres and further research is needed.

Nowadays, most of these feature extraction algorithms are implemented in state-of-the-art audio frameworks, commonly employed in most MIR studies. In this survey, we have reviewed the emotionally-relevant features from 4 common audio frameworks (Marsyas [64], MIR Toolbox (MIR TB) [65], PsySound [66] and Essentia [67]), based on the identified relations between different musical elements and emotions (as discussed in Section 3). The available frameworks vary greatly in many aspects, from user-friendliness to computational efficiency or the number of implemented algorithms. Some are aimed to research, requiring specific environments (e.g., MATLAB), while others are designed with performance in mind, more suited to be used in industry. For an in-depth review, see [68], [69].

In the following, we catalog the audio features that have been proposed in the literature over the years and are now available in these frameworks, organizing them according to the musical dimensions to which they are closest. Besides these frameworks, which implement most of the state-of-the-art audio features, in a recent work, we have contributed with a set of emotionally-relevant audio features, comprising mostly expressivity and musical texture feature [5]. As will be discussed, those features are noticeably underrepresented in the discussed audio frameworks.

Many of the features are extracted repeatedly for smaller excerpts (analysis windows) of the entire audio clip, returning series of data. These frame-level features are usually integrated using statistical moments such as mean, standard deviation, skewness and kurtosis, as well as maximum and minimum, before being used with machine learning techniques.

4.1 Melody Features

In this section we describe the audio features that capture information primarily related with melody and its components, as summarized in Table 9.

TABLE 9
Melody Features

ME	*Feature*	*Available in*
Pitch	*Pitch*	Marsyas, MIR TB, PsySound3, Essentia
	Virtual Pitch Features	PsySound3
	Pitch Salience	MIR TB, Essentia
	Predominant Melody F0	Essentia
	Pitch Content	Marsyas (unconf.)
Pitch variation	*MIDI Note Number stats*	[5]
Pitch range	*Register Distribution*	[5]
Melodic intervals	*n.a.*	*n.a.*
Melodic direction and contour	*Note Smoothness stats*	[5]
Melodic movement	*Ratios of Pitch Trans.*	[5]

4.1.1 Pitch

Pitch represents the perceived fundamental frequency of a sound. It is one of the three major auditory attributes of sounds, along with loudness and timbre. Pitch (as an audio feature) typically refers to the fundamental frequency of a monophonic sound signal and can be calculated using various techniques. One common method to calculate pitch, employed in Marsyas, MIR Toolbox and Essentia is the YIN algorithm [70]. PsySound3 also implements Swipe and Swipe′ algorithms proposed by Camacho [71].

4.1.2 Virtual Pitch Features

Ernst Terhardt *et al*. proposed an algorithm to extract virtual pitch, which is related to the psychoacoustics and modelling of the perceived pitch [72]. The PsySound3 framework implements this algorithm.

4.1.3 Pitch Salience

The perception of pitch, in particular its salience, is a complex idea that can be roughly explained as how noticeable (that is, strongly marked) is the pitch in a sound, and was proposed as a quick measure of tone sensation. Pure tones have an average pitch salience value close to 0 whereas sounds containing several harmonics in the spectrum have higher salience values. Different approaches have been proposed to extract pitch salience, e.g., [73]. This feature is present in the MIR Toolbox and Essentia.

4.1.4 Predominant Melody F0

Several authors have proposed algorithms to estimate the fundamental frequency (F0) of the predominant melody in both polyphonic and monophonic music audio signals. This is still an open research problem, and most of the audio frameworks do not include polyphonic audio melody F0 extractors. Still, some of the proposed algorithms are nowadays available as separate tools, e.g., the MELODIA algorithm [73], provided in Essentia.

4.1.5 Pitch Content

Tzanetakis proposed a set of simple features extracted from folded and unfolded pitch histograms (in the folded pitch

histogram all notes are mapped to a single octave) to describe pitch information [64]:

- FA0: Amplitude of the maximum peak of the folded histogram;
- UP0: Period of the maximum peak of the unfolded histogram;
- IPO1: Pitch interval between the two most prominent peaks of the folded histogram;
- SUM: The overall sum of the histogram.

Although the author described these features in his PhD thesis about the Marsyas framework, the current documentation seems to ignore them. Due to this we could not confirm that the framework is able to extract them.

4.1.6 MIDI Note Number (MNN) Statistics

Panda *et al.* [5] proposed 6 statistics based on the MIDI note number of each note: *MIDImean*, i.e., the average MIDI note number of all notes, *MIDIstd* (standard), *MIDIskew* (skewness), *MIDIkurt* (kurtosis), *MIDImax* (maximum) and *MIDImin* (minimum).

These features rely on the melody transcription of the original audio waveform. In that work, the authors employed the works by Salamon and Gómez [73] and Dressler [74] to estimate predominant fundamental frequencies as well as saliences. The resulting pitch trajectories are then segmented into individual MIDI notes following the work by Paiva *et al.* [75].

4.1.7 Register Distribution

This class of features proposed in [5] indicates how the notes of the predominant melody are distributed across different pitch ranges. Each instrument and voice type have different ranges, which in many cases overlap. The authors selected 6 classes, based on the vocal categories and ranges for non-classical singers. The resulting metrics are the percentage of MIDI note values in the melody 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).

In addition, the authors also propose the register distribution per second, 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.

4.1.8 Note Smoothness (NS) Statistics

Also related to the characteristics of the melody contour, Panda *et al.* [5] propose a note smoothness feature as 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.

4.1.9 Ratios of Pitch Transitions

In Panda *et al.* [5], the abovementioned 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, Transitions to Lower Pitch Notes Ratio and Transitions to Equal Pitch Notes Ratio. There, the ratio of the number of specific transitions to the total number of transitions is computed.

TABLE 10
Harmony Features

ME	*Feature*	*Available in*
Harmonic perception (harmonic intervals)	Inharmonicity	MIR TB, Essentia
	Chromagram	Marsyas, MIR TB, Essentia
	Chord Sequence	Essentia
Harmony (tonality)	Tuning Frequency	Essentia
	Key Strength	MIR TB, Essentia
	Key and Key Clarity	MIR TB, Essentia
	Tonal Centroid Vector	MIR TB
	HCDF	PsySound3
	Sharpness	PsySound3
Harmony (mode)	Modality	MIR TB, Essentia

4.2 Harmony Features

In this section we describe the audio features that capture information primarily related with harmony and its components (Table 10).

4.2.1 Inharmonicity

The inharmonicity feature is based on number of partials that are not multiples of the fundamental frequency. Inharmonicity influences the timbre perception of a given sound. One approach to compute this was proposed by Peeters *et al.* [76] and is implemented in Essentia. The MIR Toolbox measures the inharmonicity as the amount of energy outside the ideal harmonic series, which presupposes that there is only one fundamental frequency [65].

4.2.2 Chromagram

The chromagram (implemented in Marsyas, MIR Toolbox and Essentia) is used to estimate the energy distribution along pitch classes. It consists of a 12-dimension vector, one for each note, from A to G# (12 semitone pitch classes), with the respective intensities in each of these classes based on the spectral peaks of the waveform. It is also known as Harmonic Pitch Class Profile (HPCP) [65].

4.2.3 Chord Sequence

Extracting chords from an audio signal is a complex task, for which researchers have yet to propose robust solutions. The existing methods to estimate this are still experimental, based on pitch class profiles [77]. Essentia implements an algorithm based on this research, able to compute the sequence of chords in a song. Such algorithm calculates the best matching major or minor triad and outputs the result as a string (e.g., A#, Bm, G#m, C). The existing implementation is marked as experimental and requires further work before being usable.

4.2.4 Tuning Frequency

The tuning frequency (available in Essentia) is an estimation of the exact frequency (in Hz) on which a song is tuned. It is used as an intermediary step for HPCP calculation and key estimation but can also be applied for classification tasks such as western vs. non-western music [77].

4.2.5 Key Strength

Key strength (MIR Toolbox and Essentia) consists in the computation of the strength of each possible key candidate to be the key of a given song (e.g., outputting scores between 0 and 1, or -1 to 1). The algorithm is based on the cross-correlation of the chromagram [77].

4.2.6 Key and Key Clarity

These features (implemented in the MIR Toolbox and Essentia) give a broad estimation of tonal center positions and their respective clarity. This is based on peak picking in the key strength curve. There, the best key(s) is given by the peak abscissa, while the key clarity is the key strength associated with the best keys, i.e., the key ordinate [65].

4.2.7 Tonal Centroid Vector (6 dimensions)

In the MIR Toolbox, the tonal centroid is represented as a 6-dimensional feature vector. It corresponds to a projection of the chords along circles of fifths, of minor thirds and of major thirds [78]. It is based on the Harmonic Network or Tonnetz, which is a planar representation of pitch relations, where pitch classes having close harmonic relations such as fifths, major/minor thirds have smaller euclidean distances on the plane. By calculating the euclidean distance between successive analysis frames of tonal centroid vectors, the algorithm detects harmonic changes such as chord boundaries from musical audio.

4.2.8 Harmonic Change Detection Function

PsySound3 implements the Harmonic Change Detection Function (HCDF), which is a method for detecting changes in the harmonic content of musical audio signals proposed by Harte *et al.* [78]. It can be interpreted as the flux of the tonal centroid, as in the distance between the harmonic regions of successive frames [78].

4.2.9 Sharpness

Sound can be subjectively rated on a scale from dull to sharp, and sharpness algorithms attempt to model this. PsySound3 implements several algorithms [66], which are essentially weighted centroids of specific loudness.

4.2.10 Modality

Several algorithms exist to estimate modality, i.e., major vs. minor, returning either a binary label, e.g., major / minor, or a numerical value, e.g., between -1 (minor) and 1 (major) [65]. In the MIR Toolbox and Essentia, the typical strategies use the estimated strength of each key and consist of:

- the difference between the strength of the strongest major and minor keys
- the sum of all the differences between each major key and its relative minor key pair.

4.3 Rhythm Features

In this section we describe the audio features that capture information primarily related with rhythm and its components (Table 11).

TABLE 11
Rhythm Features

ME	*Feature*	*Available in*
Tempo	Beat Spectrum	MIR TB
	Beat Location	Marsyas, Essentia
	Onset Time	MIR TB, Essentia
	Event Density	MIR TB
	Average Duration of Events	MIR TB
	Tempo	Marsyas, MIR TB, Essentia
	PLP Novelty Curves	Essentia
	HWPS	Marsyas
Tempo and Note Values	Metrical Structure	MIR TB
	Metrical Centroid and Strength	MIR TB
	Note Duration statistics	[5]
	Note Duration Distribution	[5]
	Ratios of Note Duration Transitions	[5]
Rhythm Types	Rhythmic Fluctuation	MIR TB
	Tempo Change	MIR TB
	Pulse / Rhythmic Clarity	MIR TB, Essentia
Rests	n.a.	n.a.

4.3.1 Beat Spectrum

The beat spectrum (MIR Toolbox) has been proposed as a measure of acoustic self-similarity as a function of time lag. It is computed from the similarity matrix, obtained by comparing the spectral similarity between all possible pairs of frames from the original audio signal [79].

4.3.2 Beat Location

Different beat tracking algorithms have been proposed over time. These algorithms estimate the beat locations in an input signal. The Essentia framework implements several beat tracker and rhythm extractor functions, e.g., the multi-feature beat tracker, which extends the idea of measuring the level of agreement between a committee of different beat tracking algorithms in a song-by-song basis [80]. Marsyas implements IBT, a real-time/off-line tempo induction and beat tracking system based on a competing multi-agent strategy that considers parallel hypotheses regarding tempo and beats [81].

4.3.3 Onset Time

Another way of determining the tempo is based on the computation of an onset detection curve, showing the successive bursts of energy corresponding to the successive pulses [76]. Peak picking is automatically performed on the onset detection curve, to show the estimated positions of the note onsets. This feature is provided by the MIR Toolbox and Essentia. In the case of the MIR Toolbox, its onset function is able to return the onset times using any of the following options: peaks, valleys, attack phase and release phase [65].

4.3.4 Event Density

This feature (MIR Toolbox) estimates the "speed" of a song based on the average number of events in a given time window, i.e., the number of note onsets per second [65].

4.3.5 Average Duration of Events

In the MIR Toolbox, the duration of events (e.g., a note) can also be estimated from its envelope. One possible approach to estimate this was proposed by Peeters *et al.* [76]. It consists in detecting attack and release phases and measuring the time (in seconds) between them when the amplitude is at least 40 percent of the maximum.

4.3.6 Tempo

Several algorithms have been proposed to estimate tempo [19], i.e., the speed of a given musical piece, usually indicated in beats per minute (BPM). This feature, available in Marsyas, the MIR Toolbox and Essentia through different alternative algorithms, is typically estimated by detecting periodicities from the onset detection curve [65].

4.3.7 Predominant Local Pulse (PLP) Novelty Curves

Grosche and Muller introduced a mid-level representation for capturing dominant tempo and predominant local pulse even from music with weak non-percussive note onsets and strongly fluctuating tempo [82]. Essentia implements this feature. While the PLP curve does not represent high-level information such as tempo, beat level or location of onset positions, it serves as a tool that may be used for tasks such as beat tracking, tempo and meter estimation.

4.3.8 Harmonically Wrapped Peak Similarity (HWPS)

Tzanetakis described a set of rhythmic content features calculated with recourse to the Beat Histograms of a song, which proved useful for musical genre classification [64]:

- A0, A1: relative amplitude of the first (A0), and second (A1) histogram peak;
- RA: ratio of the amplitude of the second peak divided by the amplitude of the first peak;
- P1, P2: Period of the first and second peak in BPM;
- SUM: histogram sum (indication of beat strength)

Subsequently, HWPS, a feature following similar principles has been proposed and integrated into Marsyas to calculate harmonicity by taking "into account spectral information in a global manner" [83].

4.3.9 Metrical Structure

This feature provides a detailed description of the hierarchical metrical structure by detecting periodicities from the onset detection curve and tracking a broad set of metrical levels [65]. This extractor is used to calculate the meter-based tempo estimation in the MIR Toolbox.

4.3.10 Metrical Centroid and Strength

These functions provide two descriptors derived from the above metrical analysis performed in the MIR Toolbox:

- Dynamic metrical centroid: estimation of the metrical activity, based on the computation of the centroid of the selected metrical level [65];
- Dynamic metrical strength: an indicator of the clarity and strength of the pulsation. Estimates whether a "clear and strong pulsation, or even a strong metrical hierarchy is present", or if the opposite is true, where "the pulsation is somewhat hidden, unclear" [65] or a complex mix of pulsations.

4.3.11 Note Duration statistics

Panda *et al.* propose note duration statistics (the same six ones, as proposed for the melody dimension), based on the duration of each note [5].

4.3.12 Note Duration Distribution

Moreover, note duration distribution features are also proposed in [5]: Short Notes Ratio, Medium Length Notes Ratio and Long Notes Ratio. Similarly, the authors compute the note duration distribution per second, for each of the three duration classes defined.

4.3.13 Ratios of Note Duration Transitions

Finally, Panda *et al.* also propose ratios of note duration transitions [5], namely, Transitions to Longer Notes Ratio, Transitions to Shorter Notes Ratio and Transitions to Equal Length Notes Ratio.

4.3.14 Rhythmic Fluctuation

This feature (present in the MIR Toolbox) estimates the rhythm content of an audio signal. This estimation is based on spectrogram computation transformed by auditory modeling followed by spectrum estimation in each band [84], i.e., the rhythmic periodicity along auditory channels.

4.3.15 Tempo Change

An indicator of tempo change over time is estimated by computing the difference between successive values of the tempo curve in the MIR Toolbox. This feature is expressed independently from the choice of a metrical level by computing the ratio of tempo values between successive frames and is expressed in logarithmic scale (base 2) [65].

4.3.16 Pulse / Rhythmic Clarity

This feature (implemented in the MIR Toolbox and Essentia) estimates the "rhythmic clarity", an indicator of the clarity and strength found in the beats estimated by tempo estimation algorithms. Distinct heuristics exist to this estimation. The most common uses the autocorrelation curve that is computed during tempo estimation [65]. Essentia computes an approximate metric calling it beats loudness.

4.4 Dynamics Features

In this section we describe the audio features that capture information primarily related with dynamics and its components (Table 12).

4.4.1 Root-Mean-Square (RMS) Energy

The RMS energy (implemented in Marsyas, the MIR Toolbox and Essentia) is used to measure the power of a signal over a window, or global energy. This is usually computed by taking the root-mean-square (RMS) [64]. It roughly describes the loudness of a musical signal.

4.4.2 Low Energy Rate

Low energy rate (available in Marsyas and the MIR Toolbox) measures the percentage of frames with less-than-average energy [64]. This metric estimates the temporal distribution of energy, in order to understand if this energy remains constant between frames or if some frames are more contrastive than others.

4.4.3 Sound Level

This descriptor (present in PsySound3) corresponds to the power sum of the spectrum for each time window, expressed in decibel. At a higher level, when appropriately calibrated, this represents the unweighted sound pressure level of the signal in each analysis window [66].

4.4.4 Instantaneous Level, Frequency and Phase

These features (implemented in PsySound3) consist in applying a Hilbert transform to the audio waveform, resulting in three different outputs: the instantaneous level, instantaneous frequency and instantaneous phase. The instantaneous level can be regarded as the sound pressure level derived from the Hilbert transform [66].

4.4.5 Loudness

Sound loudness is the subjective perception of the intensity of a sound. This metric is measured in sones, where a doubling in sones corresponds to a doubling of loudness [66]. Several loudness metrics have been proposed over the years, which are available in PsySound3 and Essentia.

4.4.6 Timbral Width

Timbral width (PsySound3) is one of six measures of timbre proposed by Malloch in a method called loudness distribution analysis [85]. Timbral width can be regarded as "a measure of the fraction of loudness that lies outside of the loudest band, relative to the total loudness" [85].

4.4.7 Volume

Volume refers roughly to the perceived "size" of the sound, or the auditory volume of pure tones. This concept was first studied by Stevens [86] and, later on, Cabrera [87] developed a computational volume model for arbitrary spectra, which was integrated into PsySound3. In his work, Cabrera proposes two diotic volume models. The first uses a weighted ratio between the binaural loudness and sharpness, which is the specific loudness centroid on the Bark scale. A second and better performing model uses a simpler centroid to overcome limitations in the method of sharpness calculation selected by the authors [87].

4.4.8 Sound Balance

Sound balance can be estimated through the Maximum Amplitude Position to Total Envelope Length Ratio (MaxToTotal and MinToTotal), provided in the MIR Toolbox and Essentia. This is a metric to understand how much the maximum amplitude (peak) in a sound envelop is off the center. To this end, the ratio between the index of the maximum (or minimum) value of the envelope of a signal and the total length of the envelope is computed. If the peak amplitude is found close to the beginning (e.g., decrescendo sounds), this ratio will be close to 0. A value of 0.5 means that the peak is close to the middle and near 1 if at the end of the sound (e.g., crescendo sounds) [69].

TABLE 12
Dynamics Features

ME	*Feature*	*Available in*
Dynamic levels (forte, piano, etc.)	RMS Energy	Marsyas, MIR TB, Essentia
	Low Energy Rate	Marsyas, MIR TB
	Sound Level	PsySound3
	Instantaneous Level, Freq. and Phase	PsySound3
	Loudness	PsySound3, Essentia
	Timbral Width	PsySound3
	Volume	PsySound3
	Sound Balance	MIR TB, Essentia
	Note Intensity statistics	[5]
	Note Intensity Distribution	[5]
Accents and changes in dynamic levels	Ratios of Note Intensity Transitions	[5]
	Crescendo and Decrescendo metrics	[5]

4.4.9 Note Intensity statistics

Panda *et al*. compute the usual 6 statistics based on the median pitch salience of each note [5].

4.4.10 Note Intensity Distribution

In addition, Panda *et al*., 2018 propose note intensity distribution features [5]. This class of features indicates how the notes of the predominant melody are distributed across three intensity ranges, leading to the following features: Low Intensity Notes Ratio, Medium Intensity Notes Ratio and High Intensity Notes Ratio. The same features are also computed per second.

4.4.11 Ratios of Note Intensity Transitions

Panda *et al*., 2018 also propose ratios of Note Intensity Transitions: Transitions to Higher Intensity Notes Ratio, Transitions to Lower Intensity Notes Ratio and Transitions to Equal Intensity Notes Ratio [5].

4.4.12 Crescendo and Decrescendo (CD) metrics

Panda *et al*. identify notes as having crescendo or decrescendo based on the intensity difference between the first and the second half of the note [5]. From these, the authors compute the number of crescendo and decrescendo notes (per note and per second). In addition, they compute sequences of notes with increasing or decreasing intensity, computing the number of sequences for both cases (per note and per sec) and the length of crescendo sequences in notes and in seconds, using the 6 usual statistics.

TABLE 13
Tone Color (Timbre) Features

ME	*Feature*	*Available in*
Amplitude envelope	Attack/Decay Time	MIR TB, Essentia
	Attack/Decay Slope	MIR TB
	Attack/Decay Leap	MIR TB
	Zero Crossing Rate	Marsyas, MIR TB, Essentia
Spectral envelope (no. harmonics)	Spectral Flatness	Marsyas, MIR TB, Essentia
	Spectral Crest Factor	Marsyas
	Irregularity	MIR TB
	Tristimulus	Essentia
	Odd-to-even harmonic energy ratio	Essentia
Spectral characteristics (e.g., spectral centroid)	Spectral Centroid	Marsyas, MIR TB, PsySound3, Essentia
	Spectral Spread	MIR TB, PsySound3, Essentia
	Spectral Skewness	MIR TB, PsySound3, Essentia
	Spectral Kurtosis	MIR TB, PsySound3, Essentia
	Spectral Entropy	MIR TB, Essentia
	Spectral Flux	Marsyas, MIR TB, Essentia
	Spectral Rolloff	Marsyas, MIR TB, Essentia
	High-frequency Energy	MIR TB, Essentia
	Cepstrum (Real/Complex)	PsySound3
	Energy in Mel/Bark/ERB Bands	MIR TB, PsySound3, Essentia
	MFCCs	Marsyas, MIR TB, Essentia
	LPCCs	Marsyas, Essentia
	Linear Spectral Pairs	Marsyas
	Spectral Contrast	Essentia
	Roughness	MIR TB, PsySound3, Essentia
	Spectral and Tonal Dissonance	PsySound3

4.5 Tone Color Features

In this section we describe the audio features that capture information related with tone color (timbre) and its components (Table 13).

4.5.1 Attack/Decay Time

One of the aspects influencing tone color is the sound envelope, which can be divided into four parts: attack, decay, sustain and release. Several descriptors can be extracted from it, mostly related with the attack phase, i.e., from the starting point of the envelope until the amplitude peak is attained. One of these descriptors is the attack time (present in the MIR Toolbox and Essentia), which consists in the estimation of temporal duration of the various attack phases in an audio signal [76]. The MIR Toolbox is also able to compute the decay time.

4.5.2 Attack/Decay Slope

The attack slope (available in the MIR Toolbox) is another descriptor extracted from the attack phase [76]. It consists on the estimation of the average slope of the entire attack phase, since its start to the peak. The MIR Toolbox is also able to extract the same information from the decay phase, related to its decrease slope [65].

4.5.3 Attack/Decay Leap

The attack leap is a simple descriptor related to the attack phase. In the MIR Toolbox, it consists in the estimation of the amplitude difference between the beginning (bottom) and the end (peak) of the attack phase [65]. As with the previous features, the MIR Toolbox outputs a similar descriptor related with the decay phase.

4.5.4 Zero Crossing Rate (ZCR)

The Zero Crossing Rate (Marsyas, MIR Toolbox Essentia) represents the number of times the waveform changes sign in a window (crosses the x-axis). It can be used as a simple indicator of change of frequency or noisiness. As an example, heavy metal music, due to guitar distortion and heavy percussion, will tend to have much higher zero crossing values than classical music [64]. Sometimes the ZCR derivative is also computed, representing the absolute value of the window-to-window change in zero crossing rate.

4.5.5 Spectral Flatness

The spectral flatness (Marsyas, MIR Toolbox, Essentia) indicates whether the spectrum distribution is smooth or spiky, i.e., estimates to which degree the frequencies in a spectrum are uniformly distributed (noise-like) [65]. It is usually computed as the ratio between the geometric mean and the arithmetic mean [76]. Marsyas adopts a different approach, proposed in [88], calculating the spectral flatness in different spectral bands.

4.5.6 Spectral Crest Factor (SCF)

The spectral crest factor [88] is a measure of the "peakiness" of a spectrum and is inversely proportional to the spectral flatness measure. It is commonly used to distinguish noise-like from tone-like sounds due to their different spectral shapes, where noise-like sounds have lower spectral crests. In Marsyas, the SCF is computed as the ratio of the maximum and mean spectrum powers of a subband.

4.5.7 Irregularity

Irregularity, also known as spectral peaks variability, is the degree of variation of the amplitude of successive spectral peaks [65]. This feature is present in the MIR Toolbox.

4.5.8 Tristimulus

The tristimulus feature [76], implemented in Essentia, quantifies the relative energy of partial tones by three parameters that measure the energy ratio of the first partial (tristimulus1), second, third and fourth partials (tristimulus2) and the remaining (tristimulus3).

4.5.9 Odd-to-Even Harmonic Energy Ratio

The odd-to-even harmonic energy (Essentia) ratio "distinguishes sounds with predominant energy at odd harmonics (such as clarinet sounds) from other sounds with smoother spectral envelopes (such as the trumpet)" [76].

4.5.10 Spectral Moments: Centroid, Spread, Skewness, and Kurtosis

The four spectral moments (implemented in the MIR Toolbox, PsySound and Essentia) are useful measures of spectral shape [76]. The spectral centroid (also available in Marsyas) is the first moment (mean) of the magnitude spectrum of the short-time Fourier Transform (STFT).

The spectral spread represents the standard deviation of the magnitude spectrum. Thus, it is a measure of the dispersion or spread of the spectrum.

Spectral skewness is the third central moment of the magnitude spectrum and it is a measure of its symmetry.

Finally, in simple terms, spectral kurtosis, or the fourth central moment of the magnitude spectrum, captures information about existing outliers.

4.5.11 Spectral Entropy

The spectral entropy of a signal is a measure of its spectral power distribution, based on Shannon entropy [89] from the information theory field. This feature is implemented in the MIR Toolbox and Essentia.

4.5.12 Spectral Flux

Spectral flux (Marsyas, MIR Toolbox, Essentia) is a measure of the amount of spectral change in a signal, i.e., the distance between the spectra of successive frames [64]. Spectral flux has also been shown by user experiments to be an important perceptual attribute in the characterization of the timbre of musical instruments [90].

4.5.13 Spectral Rolloff

Spectral rolloff (Marsyas, MIR Toolbox, Essentia) is often used as an indicator of the skewness of the frequencies present in a window. According to Tzanetakis [64], the spectral rolloff is defined as the frequency R_t below which 85 percent of the magnitude distribution is concentrated. The percentage varies among authors, but 85 percent is the current default value for most frameworks.

4.5.14 High-Frequency Energy

Several algorithms have been proposed to estimate the high-frequency content in a signal. Brightness (also called high-frequency energy) is one of such algorithms, implemented in the MIR Toolbox. This typically consists in fixing a minimum frequency value and measuring the amount of energy above that frequency [65]. The Essentia framework implements a different algorithm, named high-frequency content (HFC), to measure the amount of high-frequency energy from the signal spectrum. HFC is computed by applying one of the several algorithms, e.g., [91].

4.5.15 Cepstrum (Real/Complex)

The cepstrum is the result of taking the inverse Fourier transform of the logarithm of the estimated spectrum of a signal [92]. It can be regarded as a measure of the rate of change in the different spectral bands. Cepstral analysis has applications in fields such as pitch analysis, echo detection and human speech processing, by providing a simple way to separate formants (due to filtering in the vocal tract) from the vocal source [93]. Cepstral analyzers are available in PsySound3.

4.5.16 Energy in Mel/Bark/ERB Bands

In audio signal processing, it is often important to decompose the original signal into a series of audio signals of different frequencies (i.e., low to high-frequency channels), enabling the study of each channel separately. This is inspired by the human cochlea, which can be regarded as a filter bank, distributing the frequencies into critical bands. Several scales have been proposed, each one using a particular range of frequencies, e.g., the Mel, Bark or Equivalent rectangular bandwidth (ERB) scales [94]. The energy in the Mel/Bark bands is computed in the MIR Toolbox and in Essentia. The energy in the ERB bands is computed in the same two frameworks, as well as PsySound3.

4.5.17 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs [95] are another measure of spectral shape. The frequency bands are positioned logarithmically on the Mel scale and cepstral coefficients are then computed based on the Discrete Cosine Transform of the log magnitude spectrum. Typically, only the first 13 cepstral coefficients are usually returned by audio frameworks. These 13 coefficients are mostly used for speech representation but Tzanetakis states that "the first five coefficients are adequate for music representation" [64]. This descriptor is provided by Marsyas, the MIR Toolbox and Essentia.

4.5.18 Linear Predictive Coding Coefficients (LPCC)

Linear predictive coding is used in speech research to represent the spectral envelope of a digital speech signal in compressed form, using to this end information of a linear predictive model [96]. LPCCs, available in Marsyas and Essentia, represent the cepstral coefficients derived from linear prediction and have been used in a wide range of speech applications, such as speech analysis, encoding and speech emotion recognition [96].

4.5.19 Linear Spectral Pairs (LSP)

Linear Spectral Pairs (available in Marsyas) are an alternative representation of linear prediction coefficients (LPC) for transmission over a channel. LSPs have several properties (e.g., smaller sensitivity to quantization noise) that make them superior to direct quantization of LPCs. Thus, LSPs are useful in speech recognition and coding [97].

4.5.20 Spectral Contrast

The octave-based spectral contrast is a feature proposed by Jiang *et al*. [98] to represent the spectral characteristics of an audio signal, specifically the relative spectral distribution. According to the authors, the feature has been tested in music type classification problems, demonstrating a "better discrimination among different music types than mel-frequency cepstral coefficients (MFCC)" [98], which is one of the features typically used in such problems. It is implemented in Essentia.

TABLE 14
Expressivity Features

ME	*Feature*	*Available in*
Articulation	Average Silence Ratio	MIR TB
	Articulation metrics	[5]
Ornamentation	Glissando metrics	[5]
	Portamento metrics	[101]
Vibrato	Vibrato metrics	[5, 101, 102]
Tremolo	Tremolo metrics	[5]

4.5.21 Roughness (Sensory Dissonance)

Sensory dissonance, also known as roughness, is related to the beating phenomenon that occurs whenever a pair of sinusoids are close in frequency [99]. This feature is implemented in Marsyas, the MIR Toolbox and Essentia using different algorithms, the method by Sethares, which estimates total roughness by averaging all dissonance estimates across all possible peak pairs of the spectrum [100].

4.5.22 Spectral and Tonal Dissonance

PsySound3 computes spectral and tonal dissonance features. Dissonance measures the harshness or roughness of the acoustic spectrum [66]. The dissonance generally implies a combination of notes that sound harsh or are unpleasant to people when played at the same time. PsySound3 provides two descriptions of acoustic dissonance: "spectral dissonance" which uses all Fourier components, and "tonal dissonance" which uses a peak extraction algorithm before calculating dissonance.

4.6 Expressivity Features

In this section we describe the audio features that capture information primarily related with expressiveness. As will be observed, we are only aware of one feature of this type in the analyzed audio frameworks. Hence, we have recently proposed a set of novel features targeting expressivity features [5]. Table 14 summarizes the available expressivity features.

4.6.1 Average Silence Ratio (ASR)

Average Silence Ratio is a feature proposed by Feng *et al.* as an estimation for articulation [3]. It is defined as the ratio of silence frames in one-second time windows. According to the author "lower ASR means fewer silence frames present in musical piece, or legato in articulation, and the higher ASR means more silence frames present in musical piece, or staccato in articulation". This feature is implemented in the MIR Toolbox.

4.6.2 Articulation Metrics

Articulation is a technique affecting the transition or continuity between notes or sounds. Panda *et al.* [5] proposed an approach to detect legato (i.e., connected notes played "smoothly") and staccato (i.e., short and detached notes). Based on their algorithm, all the transitions between notes in the song clip are classified and, from them, several metrics are extracted such as ratio of staccato, legato and other transitions and longest sequence of each articulation type.

4.6.3 Glissando Metrics

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. Panda *et al.* [5] proposed a glissando detection algorithm based on which several glissando features are extracted, e.g., glissando presence, extent, duration, direction, slope and glissando to non-glissando ratio (i.e., the ratio of notes containing glissando to the total number of notes).

4.6.4 Portamento Metrics

Computational models of portamento, the smooth and monotonic increase or decrease in pitch from one note to the next, were proposed in [101] by using Hidden Markov Models in the vibrato-free pitch curve (flatten out).

4.6.5 Vibrato Metrics

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. Panda *et al.* [5] proposed a vibrato detection algorithm based on the analysis of F0 sequence of each note, from which several features are extracted, e.g., vibrato presence, rate, extent, coverage, high-frequency coverage, vibrato to non-vibrato ratio and vibrato notes base frequency. Other approaches to extract vibrato parameters were proposed, such as using filter diagonalization methods [101] or directly from the spectrogram using predefined vibrato templates [102].

4.6.6 Tremolo Metrics

Tremolo is a trembling effect, somewhat similar to vibrato but regarding change of amplitude. Although, in the survey presented in Section 3, we have not found any relations between tremolo and emotion, we decided to extract a number of tremolo metrics, based on a tremolo detection algorithm, similar to our vibrato detection approach [5]. There, 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. Several tremolo features are extracted, e.g., tremolo presence, rate, extent, coverage, and tremolo to non-tremolo ratio.

4.7 Texture Features

In this section we describe the audio features that capture information primarily related with musical texture. To the best of our knowledge, none of the features studied or found in the analyzed audio frameworks are primarily related with musical texture. As such, we have recently proposed a set of novel musical texture features in [5], where the sequence of multiple frequency estimates was employed to measure the number of simultaneous layers in each frame of the entire audio signal, leading to the features summarized in Table 15 and described below.

4.7.1 Musical Layers Statistics

Panda *et al.* proposed musical layer statistics [5]. There, the number of multiple F0s are estimated from each frame of

TABLE 15
Texture Features

ME	Feature	Available in
Number of layers and density	Musical Layers statistics	[5]
	Musical Layers Distribution	[5]
	Ratio of Musical Layers Transitions	[5]
Texture type	n.a.	n.a.

TABLE 16
Form Features

ME	Feature	Available in
Form Complexity	Structural Change	[103]
Organization Levels	Similarity Matrix	MIR TB
	Novelty Curve	MIR TB
Song Elements	Higher-Level Form Analysis	[104-106]

the song clip. The number of layers in a frame is defined as the number of obtained multiple F0s in that frame. Then, the 6 usual statistics regarding the distribution of the number of musical layers across frames were computed.

4.7.2 Musical Layers Distribution

Additionally, in [5] 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 class is computed.

4.7.3 Ratio of Musical Layers Transitions

Panda *et al*. [5] proposed these features to capture information about the changes from a specific musical layer sequence to another. They employ the number of different fundamental frequencies 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). In addition, they also compute the length in seconds of the longest segment for each musical layer.

4.8 Form Features

In this section we describe the audio features that capture information primarily related with musical form. Extracting musical form and structure information directly from the audio signal is more difficult when compared to other lower level features (e.g., spectral/timbral statistics). Thus, few computational extractors are available today, as presented in Table 16 and described below.

4.8.1 Structural Change

The amount of change of various underlying basis features at different time intervals, combined into a meta-feature, correlates with the human perception of complexity in music [103]. The typical implementation uses chroma, rhythm and timbre information and exclusively aims at discovering the quantity of change, illustrating it with a visual audio flower plot [103].

4.8.2 Similarity Matrix

Some approaches estimate musical structure based on the similarity between adjacent segments or frames [65]. These similarities are often represented using an inter-frame or inter-segment similarity matrix, showing the differences between all possible pairs of frames from the input audio signal. The similarity matrix computation uses a specific set of frame statistics (e.g., spectral features) and a distance function, to calculate the proximity between each pair of frames. As an example, the MIR Toolbox can use MFCCs, key strength, tonal centroid, chromagram and others with one of several distance functions.

4.8.3 Novelty Curve

Based on the specific musical characteristics of each segment or frame, obtained for instance with a similarity matrix, a novelty curve can be obtained by comparing the successive frames to estimate temporal changes in the song [65]. In this novelty curve, implemented in the MIR Toolbox, the probability of transitioning to a different state over time is represented by the curve peaks.

4.8.4 Higher-Level (HL) Form Analysis

Modeling the fundamental aspects of musical sections in a unified way to identify song elements such as intro, bridge or chorus is still and open problem. Some of the most promising approaches apply higher-level solutions combining low-level features, statistics and machine learning. These include hierarchical semi-markov models [104], convex non-negative matrix factorization, spectral clustering [105] and deep learning [106].

4.9 Vocal Features

A few works have studied emotion in speaking and singing voice [107], as well as the related acoustic features [108]. 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" [15].

To this end, Panda *et al*. [5] applied the singing voice separation approach proposed by Fan *et al*. [109] (although separating the singing voice from accompaniment in an audio signal is still an open problem) and the Voice Analysis Toolkit, a "set of Matlab code for carrying out glottal source and voice quality analysis"[3] to extract the features summarized in Table 17 and described below.

4.9.1 All Features From the Vocals Channel

Besides extracting features from the original audio signal, Panda *et al*. [5] also extracted the previously described features from the signal containing only the separated voice.

4.9.2 Voice and Unvoiced Statistics

In [5], the authors also proposed statistics related to the amount of voiced and unvoiced sections in a song. These include, among others, the number of voice segments, the mean, maximum, minimum, standard deviation, kurtosis

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

TABLE 17
Vocal features

Feature	Available in
All Features from the Vocals Channel	[5]
Voiced and Unvoiced statistics	[5]
Creaky Voice statistics	[5]

and skewness of the duration of voice segments, as well as the number of voice segments per second.

4.9.3 Creaky Voice Statistics

Panda *et al.* [5] computed statistics related with the presence of creaky voice, "a phonation type involving a low frequency and often highly irregular vocal fold vibration, [which] has the potential [. . .] to indicate emotion" [110].

4.10 High-Level Features

Finally, frameworks such as the MIR Toolbox and Essentia provide a few experimental higher-level features, related with complex concepts such as emotion, genre or danceability. Most, if not all, of these are predictors, combining classification algorithms and previously gathered data to label the source audio files into a fixed set of tags. A summary of these predictors is presented in Table 18 and listed below.

4.10.1 Emotion

The MIR Toolbox extracts an emotion descriptor based on the analysis of the audio content of a given recording. The output is given in two distinct paradigms: a categorical approach comprising 5 emotions and a 3-dimensional space composed of activity (energetic arousal), valence (pleasure-displeasure continuum) and tension (tense arousal).

The classification process is based on the work by Eerola *et al.* [111] and uses multiple linear regression with the 5 best performing predictors. Given its reliance on previously established weights, this extractor is only reliable in the MIR Toolbox version (v1.3) where it was initially "calibrated". Newer versions output "distorted results" [65].

The Essentia library implements a similar feature, classifying songs in 4 distinct emotions. It contains pre-trained models and requires the Gaia library to apply similarity measures and classifications on the extracted features [67].

4.10.2 Classification-Based Features (Genre, etc.)

In a similar way to the emotion descriptor extractor (or predictor), Essentia also includes Gaia trained models for [67]:

- musical genre (using 4 different databases)
- ballroom music classification
- western / non-western music
- tonal / atonal
- danceability
- voice / instrumental
- gender (male / female singer)
- timbre classification

These musical descriptors work as a typical classification problem, by extracting a set of features from the source audio signals and feeding them to classification models trained with them in other datasets.

TABLE 18
High-Level Features

Feature	Available in
Emotion	MIR Toolbox, Essentia
Classification-based Feat. (genre, etc.)	Essentia
Danceability	Essentia
Dynamic Complexity	Essentia

The genre feature is particularly relevant for music emotion recognition since some emotions are frequently associated with specific genres, as concluded by Laurier [4]. The author used automatic genre classification to improve his previous emotion classification results.

4.10.3 Danceability

As opposed to the aforementioned danceability extractor built as a pre-trained classification model, Streich proposed a low-level audio feature derived from Detrended Fluctuation Analysis (DFA) to characterize audio signals in terms of its danceability [112].

4.10.4 Dynamic Complexity

Streich also studied the automated estimation of the complexity of music based on the musical audio signal, proposing a set of complexity descriptors [112]. The proposed algorithms focus on aspects of acoustics, rhythm, timbre, and tonality. The Essentia library implements an extractor to estimate dynamic complexity, or whether a song contains a high dynamic range. This descriptor consists in the average absolute deviation from the global loudness level estimate on the dB scale.

5 Discussion and Research Directions

5.1 Feature Analysis Along Musical Dimensions

Table 19 presents the number of described features per musical dimension.

As abovementioned, many of these features are frame-level features, which are normally integrated using statistical moments. This increases the final number of descriptors to several hundred [5] and is especially true for tone color features, where some features divide the audio signal in bands and output time-series data (e.g., MFCCs). As such, and based on the figures in Table 19, we conclude that the number of available audio features is very unbalanced

TABLE 19
Number of Audio Descriptors Per Musical Dimension

Musical dimension	Number of features	Percentage of total
Melody	9	10.6%
Harmony	10	11.8%
Rhythm	16	18.8%
Dynamics	12	14.1%
Tone Color	25	29.4%
Expressivity	6	7.1%
Texture	3	3.5%
Form	4	4.7%
Total	*85*	*100%*

across musical dimensions. Musical texture, expressivity and form are especially lacking, in contrast to tone color, which is the most represented category, mostly due to the large set of spectral features available (centroid, etc.). In [5], we have contributed to reduce that imbalance by proposing emotionally-relevant features, particularly for the expressivity and texture dimensions.

The low number of texture, form and expressivity features is not a surprise. We believe this is caused by two main reasons: i) on the one hand, the difficulty to create robust algorithms to capture such music elements; ii) on the other hand, the lack of music psychology studies on the relations between emotion and those dimensions, which could drive the creation of computational models.

Regarding the analysis of the importance of specific features to emotion recognition, few studies have addressed this issue in a systematic way, e.g., [5]. There, the conducted analysis, based on Russell's emotion quadrants [28], suggested that tone color features (particularly spectral features) dominated all quadrants, possibility due to their prevalence (as discussed above). Nevertheless, texture features were in the top5 for quadrant 2 (anxiety quadrant, or Q2) and proved relevant for Q1 (happiness), as well, helping to improve the classification performance of the proposed algorithm. Vibrato was also an important feature for Q2. As for Q3 (depression), besides tonal features, texture, inharmonicity and tremolo also proved relevant, along with vocal features. Finally, dynamics, texture and expressivity features (namely, vibrato) were important to discriminate Q4 (contentment).

Besides the lack of texture, form and expressivity features, "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" [5]. Thus, in the next section we discuss research directions to advance the state-of-the-art in the creation of novel emotionally-relevant features for each musical dimension.

5.2 Novel Audio Futures: Research Directions

5.2.1 Form

Regarding computational models of form complexity, we are only aware of one work, which might work as a surrogate of musical complexity [103]. Higher-level features to capture form types from audio are still missing and some recent works have been attacking the problem with higher level solutions, e.g., employing machine learning to identify elements such as verse and chorus [104], [105], [106].

The impact of other elements of form on emotion, e.g., organizational levels (passage, piece, cycle) or song elements (introduction, chorus, bridges, etc.), should be further researched by the music psychology community, despite a few computational models found in the literature that might partially capture such information (e.g., similarity matrix and novelty curve).

5.2.2 Texture

The texture dimension, as abovementioned, requires further music psychology studies to better understand how it influences emotion. Nevertheless, the features we proposed in [5] proved relevant, namely the number of musical layers in the recognition of happy music.

These features only approximate the actual number of layers in a song, hence more advanced computational models are needed, probably requiring robust source separation and instrument recognition in polyphonic music. This is an active research problem (e.g., [113]), with great advances in the last years due to the application of deep learning models, as is the case of the Spleeter library, able to perform various types of separation (e.g., vocals, accompaniment, drums, bass, and others) [114].

Tackling this problem would also serve the creation of algorithms for the detection of texture types (monophonic, homophonic, polyphonic) and density (thin, thick), for which no computational models are known (see Table 15).

5.2.3 Expressivity

Regarding expressivity, the music psychology community has offered important inputs to understand its impact on emotion. Yet, despite our contributions with several articulation (staccato and legato), glissando, vibrato and tremolo metrics, this dimension still lacks computational models, particularly for the detection of ornamentations other than glissando and portamento (see Tables 6 and 14). Also, the algorithms we proposed were only indirectly evaluated through their impact on emotion classification, and so ground truth data on those problems is needed.

5.2.4 Melody

As for the other musical dimensions, music psychology researchers have provided a great amount of knowledge that could be further exploited to create computational models that capture such musical elements.

Starting with melody, most melodic elements are reasonably covered, as summarized in Table 9. However, features for melodic intervals are still missing. Moreover, further computational features related to melodic movement, direction and contour should be developed. As with many other problems in Music Information Retrieval, problems such as full or melody transcription are still open, which limits the accuracy of current MER systems that rely on them. This also applies to computational models of the dimensions discussed below (e.g., tonality and rhythm).

5.2.5 Harmony

As for harmony, all elements with emotional relevance have computational features to capture them (Table 10): harmonic perception (e.g., inharmonicity), tonality (e.g., tonal centroid vector) and mode (e.g., modality).

5.2.6 Rhythm

Regarding rhythm, although most rhythmic elements are reasonably covered this dimension is missing computational features that capture rest characteristics (Table 11). Still, higher-level audio features that capture the types of rhythm (regular, irregular, complex, fluent, etc.) are still missing (see Tables 3 and 11).

5.2.7 Dynamics

As for dynamics, all elements have associated features (Table 12). Still, computational models to detect the types of dynamic levels (forte, piano, etc.) would be beneficial.

5.2.8 Tone Color

The tone color dimension is also reasonably well covered, particularly regarding spectral characteristics (see Table 13). Still, as with musical texture, tone color would also benefit from accurate instrument recognition in polyphonic context. Moreover, this dimension would also benefit from higher-level features on the types of amplitude envelope (e.g., round, sharp).

5.2.9 Vocal Features

As for vocal features, with the recent advances in areas such as source separation, as previously described, new paths should be explored. For instance, additional features that proved useful for speech emotion should be taken into consideration [16]. Moreover, the idea can be extended, e.g., by further separating the accompaniment and analyzing each layer in isolation, since they may sometimes carry different emotional information [15]. This can be complemented with genre or even lyrical information (natural language processing) and integrated with a meta-classifier.

5.3 Deep Learning Perspectives

Finally, besides the classical handcrafted feature engineering approach, deep learning/feature learning techniques have attracted great attention in the last years. The most notable example is the resurgence of neural network techniques, specifically deep learning, to a myriad of problems, fueled by the improvements in computer processing (e.g., using graphic processing units). Several MER studies have already employed techniques such as convolutional and recurrent neural networks [10].

Despite (so far) slight improvements in classification accuracy, such approaches raise several points that must be considered. First, to fully exploit the potential of deep learning solutions, massive amounts of good quality data are required. Unfortunately, the creation of large MER datasets have been known to be problematic due to the associated subjectivity and complexity of data collection [5]. Hence, strategies to obtain sizeable and good quality data for audio MER are a key need.

Also, deep learning models are opaque in the sense that the extracted features are often difficult to interpret, which hinders the possibility to acquire novel knowledge regarding the relations between emotions and the extracted features. In fact, "although deep neural networks have exhibited superior performance in various tasks, interpretability is always [the] Achilles' heel" of such approaches, despite a few efforts to address it, as surveyed in [115]. Hence, interpretability issues in deep neural networks are another important problem to tackle in the future.

5.4 Audio-Based Symbolic Features

As discussed, some approaches bridge the audio and the symbolic MER domains by integrating an audio transcription step into the feature extraction stage. Hence, the approached followed in [5] can be further exploited by integrating symbolic (MIDI) features available in several frameworks, e.g., MIDI Toolbox or jSymbolic [2].

6 Conclusion

This article offered a comprehensive review of the current emotionally-relevant computational audio features. Unlike previous broad MER surveys, this review offered a deep and specific review on that key MER problem.

This survey was supported by the music psychology literature on the relations between eight musical dimensions (melody, harmony, rhythm, dynamics, tone color, expressivity, texture and form) and specific emotions. From this review, we concluded that computational audio features able to capture elements of musical form, texture and expressivity are especially needed to break the current glass ceiling in MER, as shown in [5]. Moreover, the development of such computational tools would benefit from further music psychology studies, particularly regarding the actual impact of musical form and texture on emotion. We believe this article opens several research lines to expand the state-of-the-art on Music Emotion Recognition.

Acknowledgments

This work was supported by the MERGE project financed by Fundação para Ciência e a Tecnologia (FCT) - Portugal.

References

[1] D. Huron, "Perceptual and cognitive applications in music information retrieval," *Cognition*, vol. 14, no. 10, pp. 83–92, 2000.
[2] 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.
[3] 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*, 2003, vol. 2, no. 2, pp. 375–376.
[4] C. Laurier, "Automatic classification of musical mood by content-based analysis," PhD Thesis, Dept. Inform. Commun. Technol., Pompeu Fabra Univ., 2011.
[5] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2018.2820691.
[6] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. 4th Int. Soc. Music Inf. Retrieval Conf.*, 2003, pp. 1–2.
[7] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning,". in *Proc. 22th ACM Int. Conf. Multimedia*, 2014, pp. 117–126.
[8] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio Speech Language Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.
[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.*, vol. 9, no. 2, pp. 240–254, Second Quarter 2018.
[10] M. Malik *et al.*, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Proc. Sound Music Comp.*, 2017, pp. 208–213.
[11] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS One*, vol. 12, no. 3, 2017, Art. no. e0173392.
[12] Ò. Celma, P. Herrera, and X. Serra, "Bridging the music semantic gap," in *Proc. Workshop Mastering Gap: From Inf. Extraction Semantic Representation*, 2006.
[13] R. Scholz, G. Ramalho, and G. Cabral, "Cross task study on MIREX recent results: An index for evolution measurement and some stagnation hypotheses," in *Proc. 17th Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 372–378.
[14] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[15] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Syst.*, vol. 24, pp. 365–389, 2018.

[16] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, 2012, Art. no. 40.

[17] A. Huq, J. P. Bello, and R. Rowe, "Automated music emotion recognition: A systematic evaluation," *J. New Music Res.*, vol. 39, no. 3, pp. 227–244, 2010.

[18] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *J. Acoust. Soc. America*, vol. 136, no. 4, pp. 1951–1963, 2014.

[19] A. Elowsson and A. Friberg, "Tempo estimation by modelling perceptual speed," Music Inf. Retrieval Eval. Exchange, 2013. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2013/EF1.pdf

[20] D. Cooke, in *The Language of Music*. London, UK: Oxford Univ. Press, 1959.

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

[22] L. B. Meyer, *Explaining Music: Essays Explorations*. Berkeley, CA: Univ. of California Press, 1973.

[23] H. Owen, in *Music Theory Resource Book*. London, UK: Oxford Univ. Press, 2000.

[24] W. Dace, "The concept of "Rasa" in sanskrit dramatic theory," *Educ. Theatre J.*, vol. 15, no. 3, pp. 249–254, 1963.

[25] Plato, "Republic III," in *Plato in Twelve Volumes*, vol. vols. 5-6. Cambridge, MA, USA: Harvard Univ. Press, (375 B.C.), 1969.

[26] Aristotle, "Politics". *Aristotle 23 Volumes*, vol. 21. Cambridge, MA, USA: Harvard Univ. Press, (IV c B.C.), 1944.

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

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

[29] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," *Music Emotion*. London, UK: Oxford Univ. Press, vol. 8, 2001, pp. 223–248.

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

[31] L.-L. Balkwill and W. F. Thompson, "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," *Music Percept.*, vol. 17, no. 1, pp. 43–64, 1999.

[32] A. Gabrielsson and E. Lindström, "The role of structure in the musical expression of emotions," in *Handbook of Music and Emotion: Theory, Research, Applications*. P.N. Juslin and J.A. Sloboda, eds., London, UK: Oxford Univ. Press, pp. 367–400, 2011.

[33] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *J. New Music Res.*, vol. 33, no. 3, pp. 217–238, 2004.

[34] T. F. Maher and D. E. Berlyne, "Verbal and exploratory responses to melodic musical intervals," *Psychol. Music*, vol. 10, no. 1, pp. 11–27, 1982.

[35] W. F. Thompson and B. Robitaille, "Can composers express emotions through music?," *Empir. Stud. Arts*, vol. 10, no. 1, pp. 79–89, 1992.

[36] N. D. Cook and T. X. Fujisawa, "The psychophysics of harmony perception: Harmony is a three-tone phenomenon," *Empir. Musicology Rev.*, vol. 1, no. 2, pp. 106–126, 2006.

[37] L. Gagnon and I. Peretz, "Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies," *Cogn. Emotion*, vol. 17, no. 1, pp. 25–40, 2003.

[38] P. N. Juslin, "Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy," *Music Sci.*, vol. 1, no. 2, pp. 225–256, 1997.

[39] A. Fernández-Sotos, A. Fernández-Caballero, and J. M. Latorre, "Influence of tempo and rhythmic unit in musical emotion regulation," *Front. Comp. Neurosci.*, vol. 10, 2016, Art. no. 80.

[40] M. Plewa and B. Kostek, "A study on correlation between tempo and mood of music," in *Proc. 133th Audio Eng. Soc. Conv.*, 2012.

[41] P. N. Juslin and R. Timmers, "Expression and communication of emotion in music performance," in *Handbook of Music and Emotion: Theory, Research, Applications*, P. N. Juslin and J. A. Sloboda, eds., London, UK: Oxford Univ. Press, pp. 452–489, 2011.

[42] G. Ilie and W. F. Thompson, "A comparison of acoustic cues in music and speech for three dimensions of affect," *Music Percept.*, vol. 23, no. 4, pp. 319–330, 2006.

[43] K. B. Watson, "The nature and measurement of musical meanings," *Psychol. Monographs*, vol. 54, pp. i–43, 1942.

[44] S. K. Langer, in *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art*. Cambridge, MA, USA: Harvard Univ Press, 1957.

[45] B. Wu, A. Horner, and C. Lee, "The correspondence of music emotion and timbre in sustained musical instrument sounds," *J. Audio Eng. Soc.*, vol. 62, no. 10, pp. 663–675, 2014.

[46] J. C. Hailstone, R. Omar, S. M. D. Henley, C. Frost, M. G. Kenward, and J. D. Warren, "It's not what you play, it's how you play it: Timbre affects perception of emotion in music," *Quart. J. Exp. Psychol.*, vol. 62, no. 11, pp. 2141–2155, 2009.

[47] B. Wu, A. Horner, and C. Lee, "Musical timbre and emotion: The identification of salient timbral features in sustained musical instrument tones equalized in attack time and spectral centroid," in *Proc. 40th Int. Comput. Music Conf.*, 2014, pp. 928–934.

[48] 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, 2015.

[49] E. Lindström, "Expression in music: Interaction between performance and melodic structure," in *Proc. Meeting Soc. Music Percept. Cogn.*, 1999.

[50] R. Timmers and R. Ashley, "Emotional ornamentation in performances of a handel sonata," *Music Percept.*, vol. 25, no. 2, pp. 117–134, 2007.

[51] M. P. Kastner and R. G. Crowder, "Perception of the major/minor distinction: IV. Emotional connotations in young children," *Music Percept.*, vol. 8, no. 2, pp. 189–201, 1990.

[52] G. D. Webster and C. G. Weir, "Emotional responses to music: Interactive effects of mode, texture, and tempo," *Motivation Emotion*, vol. 29, no. 1, pp. 19–39, 2005.

[53] A. H. Gregory, L. Worrall, and A. Sarge, "The development of emotional responses to music in young children," *Motivation Emotion*, vol. 20, no. 4, pp. 341–348, 1996.

[54] R. McCulloch, "Modality and children's affective responses to music," *Undergraduate project for the Perception and Performance course (Ian Cross, instructor)*, 1999. [Online]. Available: www.ms.cam.ac.uk/~ic108/PandP/McCulloch99/McCulloch99.html

[55] Y. Broze, B. T. Paul, E. T. Allen, and K. M. Guarna, "Polyphonic voice multiplicity, numerosity, and musical emotion perception," *Music Percept.*, vol. 32, no. 2, pp. 143–159, 2014.

[56] M. Imberty, *Understanding Music: Phychological Music Semantics (Entendre la Musique: Sémantique Psychologique de la Musique)*. Dunod, 1979. [Online]. Available: https://www.amazon.com/Entendre-musique-Se%CC%81mantique-psychologique-Psychismes/dp/204010920X

[57] V. J. Konečni and M. P. Karno, "Empirical investigations of the hedonic and emotional effects of musical structure," *Musikpsychologie*, vol. 11, pp. 119–137, 1994.

[58] B. Tillmann and E. Bigand, "Does formal musical structure affect perception of musical expressiveness?," *Psychol. Music*, vol. 24, no. 1, pp. 3–17, 1996.

[59] A. Gabrielsson and P. N. Juslin, "Emotional expression in music," in *Handbook of Affective Sciences (Series in Affective Science)*. R. J. Davidson eds., Oxford Univ. Press, London, UK, 2003, pp. 503–534.

[60] E. Schubert, "Measurement and time series analysis of emotion in music," PhD Diss., School of Music and Music Education, Univ. of New South Wales, 1999.

[61] D. Huron, "What is a musical feature? Forte's analysis of brahms's opus vol. 51, no. 1, revisited," *Online J. Soc. Music Theory*, vol. 7, no. 4, 2001.

[62] A. Rodà, S. Canazza, and G. De Poli, "Clustering affective qualities of classical music: Beyond the valence-arousal plane," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 364–376, Fourth Quarter 2014.

[63] M. Schedl, E. Gomez, E. S. Trent, M. Tkalcic, H. Eghbal-Zadeh, and A. Martorell, "On the interrelation between listener characteristics and the perception of emotions in classical orchestra music," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 507–525, Fourth Quarter 2018.

[64] G. Tzanetakis, "Manipulation, analysis and retrieval systems for audio signals," PhD diss., Dept. Comput. Sci., Princeton Univ., 2002.

[65] O. Lartillot, in *MIR Toolbox 1.7.1 User's Manual*. Oslo, Norway: Univ. Oslo, 2018.

[66] 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.
[67] D. Bogdanov *et al.*, "ESSENTIA: An audio analysis library for music information retrieval," in *Proc. 14th Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 493–498.
[68] D. Moffat, D. Ronan, and J. D. Reiss, "An evaluation of audio feature extraction toolboxes," in *Proc. 18th Int. Conf. Digit. Audio Effects*, 2015, pp. DAFX-1–DAFX-7.
[69] R. Panda, "Emotion-based analysis and classification of audio music," PhD diss., Dept. Inform. Eng., Univ. of Coimbra, 2019.
[70] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. America*, vol. 111, no. 4, pp. 1917–1930, 2002.
[71] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator," PhD Diss., Univ. of Florida, 2007.
[72] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. America*, vol. 71, no. 3, pp. 679–688, 1982.
[73] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 20, no. 6, pp. 1759–1770, Apr. 2012.
[74] K. Dressler, "Automatic transcription of the melody from polyphonic music," PhD Diss., Faculty Elect. Eng. Inform. Technol., Ilmenau Univ. of Technol., 2016.
[75] R. P. Paiva, T. Mendes, and A. Cardoso, "From pitches to notes: Creation and segmentation of pitch tracks for melody detection in polyphonic audio," *J. New Music Res.*, vol. 37, no. 3, pp. 185–205, 2008.
[76] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. America*, vol. 130, no. 5, pp. 2902–2916, 2011.
[77] E. Gómez, "Tonal description of music audio signals," PhD Thesis, Dept. Technol., Pompeu Fabra Univ., 2006.
[78] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. 1st ACM Workshop Audio Music Comput. Multimedia*, 2006, pp. 21–26.
[79] J. T. Foote, M. L. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. 3rd Int. Conf. Music Inf. Retrival*, 2002, pp. 265–266.
[80] J. Zapata, M. E. P. Davies, and E. Gómez, "Multi-feature beat tracking," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 4, pp. 816–825, Apr. 2014.
[81] J. L. Oliveira, F. Gouyon, L. G. Martins, and L. P. Reis, "IBT: A real-time tempo and beat tracking system," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp, 291–296.
[82] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. 10th Int. Soc. for Music Inf. Retr. Conf.*, 2009, pp. 189–194.
[83] M. Lagrange, L. G. Martins, and G. Tzanetakis, "A computationally efficient scheme for dominant harmonic source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 165–168.
[84] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 570–579.
[85] S. N. Malloch, "Timbre and technology: An analytical partnership," PhD Diss., Univ. of Edinburgh, 1997.
[86] S. S. Stevens, "The volume and intensity of tones," *Amer. J. Psychol.*, vol. 46, no. 3, pp. 397–408, 1934.
[87] D. Cabrera, "The size of sound: Auditory volume reassessed, " in *Proc. Australas. Comput. Music Assoc. Conf.*, 1999.
[88] E. Allamanche, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer, "Content-based identification of audio material using MPEG-7 low level description," in *Proc. 2nd Int. Symp. Music Inf. Retrieval*, 2001.
[89] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[90] J. M. Grey, "An exploration of musical timbre," PhD Diss., Dept. Psychology, Stanford Univ., 1975.
[91] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proc. Int. Comp. Music Conf.*, 1996.
[92] B. P. Bogert, J. R. Healy, and J. W Tukey, "The quefrency alanysis (SIC) of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proc. Symp. Time Series Anal.*, 1963, pp. 209–243.
[93] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. America*, vol. 41, pp. 293–309, 1967.
[94] J. Harrington and S. Cassidy, in *Techniques in Speech Acoustics*. The Netherlands: Kluwer Academic Publishers, 1999.
[95] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech*, vol. TASSP-28, no. 4, pp. 357–366, Aug. 1980.
[96] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
[97] F. Zheng, Z. Song, L. Li, W. Yu, and W. Wu, "The distance measure for line spectrum pairs applied to speech recognition," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, 1998, pp. 1123–1126.
[98] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proc IEEE Int. Conf. Multimedia Expo*, 2002, pp. 113–116.
[99] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *J. Acoust. Soc. America*, vol. 38, no. 4, pp. 548–560, 1965.
[100] W. A. Sethares, in *Tuning, Timbre, Spectrum, Scale*. Berlin, Germany: Springer, 1998.
[101] L. Yang, K. Z. Rajab, and E. Chew, "AVA: An interactive system for visual and quantitative analyses of vibrato and portamento performance styles," in *Proc. 17th Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 108–114.
[102] J. Driedger, S. Balke, S. Ewert, and M. Müller, "Template-based vibrato analysis of music signals," in *Proc. 17th Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 239–245.
[103] M. Mauch and M. Levy, "Structural change on multiple time scales as a correlate of musical complexity," in *Proc. 12th Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 489–494.
[104] G. Shibata, R. Nishikimi, E. Nakamura, and K. Yoshii, "Statistical music structure analysis based on a Homogeneity-, Repetitiveness-, and regularity-aware hierarchical hidden semi-markov model," in *Proc. 20th Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 268–275.
[105] B. McFee and D. P. W. Ellis, "Analyzing song structure with spectral clustering," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 405–410.
[106] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 417–422.
[107] 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, 2015.
[108] 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, 2015, Art. no. 19.
[109] 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.
[110] A. Cullen, J. Kane, T. Drugman, and N. Harte, "Creaky voice and the classification of affect," in *Proc. Workshop Affect. Soc. Speech Signals*, 2013.
[111] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 621–626.
[112] S. Streich, "Music complexity: A multi-faceted description of audio content," PhD Diss., Dept. Technol., Pompeu Fabra Univ., 2007.
[113] S. Gurunani, M. Sharma, A. Lerch, "An attention mechanism for musical instrument recognition," in *Proc. 20th Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 83–90.
[114] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, 2020, Art. no. 2154.
[115] Q. Zhang and S. Zhu, "Visual interpretability for deep learning: A survey," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.

Renato Panda received the bachelor's, master's, and PhD degrees from the University of Coimbra. He currently is an invited professor with the Polytechnic Institute of Tomar. He is a member of the Cognitive and Media Systems group at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests include music emotion recognition (MER) and music information retrieval (MIR).

Ricardo Malheiro received the bachelor's degree in mathematics, and the master's degree in informatics engineering (Licenciatura - 5 years), and PhD degree, all from the University of Coimbra. He is currently a professor with Miguel Torga Higher Institute, Coimbra. He is also a member of the CMS research group at CISUC. His main research interests include natural language processing, detection of emotions in music lyrics and text and text/data mining.

Rui Pedro Paiva received the bachelor's degree in 1996, master's degree in 1999, and the doctoral degree, in 2007, all from the Department of Informatics Engineering of the University of Coimbra. He is currently a professor with the Department of Informatics Engineering with the University of Coimbra. He is also a member of the CMS group at CISUC. His main research interests include the areas of MIR and health informatics. The common research interests include the study of feature engineering, machine learning and signal processing to the analysis of musical and bio signals