Music Emotion Recognition: toward new, robust standards in personalized and context-sensitive applications*

Juan Sebastián Gómez-Cañón, Estefanía Cano, Tuomas Eerola,
Perfecto Herrera, Xiao Hu, Yi-Hsuan Yang, Emilia Gómez

1 Introduction

Emotion is one of the main reasons why people engage and interact with music [1] – songs can express our inner feelings, produce goosebumps, bring us to tears, share an emotional state with a composer or performer, or trigger specific memories. An interest for a deeper understanding of the relation between music and emotion has motivated researchers from various areas of knowledge for decades [2], including computational researchers. Imagine an algorithm that could "predict" the emotions that a listener perceives in a musical piece or one that could dynamically generate music adapting to the mood of a conversation in a film – a particularly fascinating and provocative idea. These algorithms typify Music Emotion Recognition (MER), a computational task that attempts to automatically recognize either the emotional content in music or the emotions induced by music to the listener [3]. To do so, emotionally-relevant features are extracted from music, processed, evaluated, and then associated with certain emotions. MER is one of the most challenging high-level music description problems in Music Information Retrieval (MIR), an interdisciplinary research field that focuses on the development of computational systems to help humans better understand music collections. MIR integrates concepts and methodologies from several disciplines including music theory, music psychology, neuroscience, signal processing, and machine learning.

 Printed publication: https://www.doi.org/10.1109/MSP.2021.3106232

Schedl et al. [4] highlighted the importance of incorporating different factors that can influence the music listening experience into user-aware music retrieval systems. The authors propose four categories of such factors: (1) *music content* – descriptors inferred from the audio signal (e.g., melody, rhythm, harmony, loudness), (2) *music context* – factors directly related to music that cannot be extracted from its content (e.g., song metadata, artist's biographies, album covers), (3) *user context* – dynamic aspects from the listener that fluctuate frequently (e.g., listening mood, uses of music, physiological signals), and (4) *user properties* – factors from the listener that are more constant (e.g., demographics, musical expertise, preference). In this light, MER systems – initially built upon classical signal processing approaches and mainly focusing on music content – can also account for individual and contextual differences depending on the application scenarios. *Context-sensitive* MER systems draw on factors from the user context, and *personalized* MER systems incorporate information from the user properties.

1.1 Perceived and induced emotions

An important distinction in MER research is that between *perceived* and *induced* emotions. Namely, a song may be *perceived* to be "sad" due to its slow tempo or its lyrics, but *induce* "joy" by triggering a specific happy memory in the listener. To further clarify this distinction, we propose the following exercise: hum the song Happy Birthday and then remember the last time you heard it. This song is most likely perceived as "happy" due to musical features such as tempo or the fact that it is written in a major key (commonly regarded as "happy" in Western music traditions). However, given a certain individual context, this song might make someone feel "sadness" if, for example, it brings a memory of a person that is not with her/him anymore. Music brings together major subtleties that magnify the complexities for its computational analysis. Historically, MER has primarily addressed the computational modeling of *perceived emotions* since it can be considered less influenced by contextual factors [3]. Conversely, research on *induced emotions* appears to be more subjective and more influenced by user context. However, given its potential for emotion regulation applications, research on induced emotions is an essential topic for the future of the field.

This initial distinction highlights the key dependence of MER research on music psychology and promotes interdisciplinary efforts in order to involve methodologies that

address the inherent subjectivity of the task, generate enriched datasets with multiple and high quality annotations, and propose positive use cases for these technologies. Nonetheless, several MER methodologies have been subject to criticism in the literature.

1.2 Limitations and criticisms of MER research

Traditional MER systems have been mainly inspired on supervised learning, relying on the existence of an annotated music emotion dataset. Indeed, the research community has openly dissected several issues of the MER field: Sturm [5] pointed out the deceptive simplicity of assembling emotion datasets from "ground truths" that are difficult to generate; Schedl et al. [6] reported low statistical inter-rater agreement of perceived emotion annotations; Lange and Frieler [7] described generalized inconsistency of subjective ratings of emotional attributes in music; Juslin [2] remarked the generalized confusion of listeners between the concepts of perceived and induced emotions, possibly impacting annotation reliability; Beveridge and Knox [8] highlighted the difficulty of discovering acoustic features responsible for expressing or inducing emotions; Schuller [9] described a paradigm shift from the design of hand-crafted features to data-learned features which has also extended to MER, where the "black box" nature of machine learning models is even more problematic for model explainability [10]. In a nutshell, these issues must be addressed in order to improve the quality and significance of MER research.

1.3 Motivation and Goals

Traditional applications for MER systems consider information retrieval use cases (i.e., organizing, managing, and recommending purposes): the automatic categorization of music pieces/collections based on emotion, and emotion-aware music recommendation for the gaming and film industries [3]. Novel personalized and context-sensitive applications can further broaden the scope where MER algorithms might aid in automatic music selection and recommendation for personal listening experiences, while allowing the potential use of music for learning and well-being [11].

The aim of this paper is to introduce readers to past and current challenges in MER, point out common pitfalls that may frustrate interdisciplinary endeavors, and guide them towards standardized and robust methodologies to face them. In section 2, we discuss current methodologies used for MER task – we review the different components of

the traditional MER framework, propose an updated emotion conceptualization framework drawn from music psychology, and summarize the main approaches for the MER task found in the literature. In section 3, we argue that robust methodological standards are necessary to tackle the following challenges of MER systems, in the interest of building personalized and context-sensitive applications: (1) open data and experimental reproducibility – we should not only release open music datasets and reproducible methodologies, but also open anonymized user data; (2) subjectivity of concepts and annotations – we should not only create better and more reliable datasets, but ones which are more relevant to a particular context and enriched with multiple annotations; (3) model explainability and interpretability – we should use computational methods that allow easier comprehension of data-driven decisions posterior to emotion prediction; (4) cultural and contextual relevance – we need to acknowledge the inherent bias towards Western music and annotators, attempting to improve cross-cultural research; and (5) ethical implications for MER applications – we must acknowledge the potential impact of MER on the listener's well-being, in terms of fairness, privacy, and social good.

2 Current Music Emotion Recognition framework

The traditional MER workflow is summarized in Figure 1 and is composed of four main blocks: (1) taxonomy definition – a particular annotation scheme is selected grounded on music theory and cognition studies for emotion modeling, (2) dataset creation – human subjects annotate perceived or induced emotions after listening to short music excerpts to define a "ground truth", (3) feature extraction – signal processing methodologies are used to extract emotionally-relevant features, which are then matched with the subjective annotations provided by humans, and (4) evaluation – a machine learning model trains on a portion of the annotated dataset and tests with the remaining fraction in order to evaluate the performance of the system. It is important to note that in this framework, factors related to user properties or user context are often not considered.

2.1 Taxonomies and musical properties of emotion – perspectives from cognition

The selection of the emotion taxonomy used to represent musically-related emotions is the initial step for the design of MER models – the chosen taxonomy is used to anno-

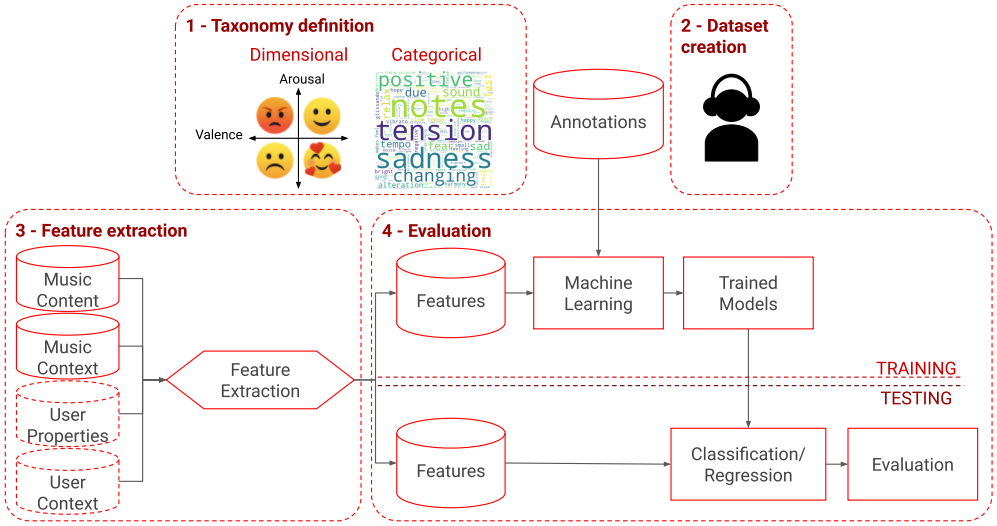

Figure 1: Traditional Music Emotion Recognition systems.

tate the music excerpts. Two predominant taxonomies are common for the collection of annotations in the MER literature (refer to [3, 12] for a comprehensive comparison): (1) the categorical/discrete approach represents distinct classes (e.g., happy, sad) [13]; (2) the dimensional/continuous approach conceptualizes emotions into specific dimensions of arousal and valence, where arousal refers to energy or activation and valence relates to pleasantness or positiveness of an emotion [14]. Each approach has advantages but also presents particular drawbacks: the categorical/discrete approach has poor resolution compared to the richness of human emotion and is naturally ambiguous by using language as an emotion descriptor [3]; the dimensional/continuous approach hinders the mapping of complex emotions (e.g., nostalgia) to a particular numeric value of arousal or valence, and has been argued to be inappropriate for music since emotions appear to fall into prototypes/categories in several situations [2]. The choice of these taxonomies is the most crucial decision for dataset creation and defines the fundamental trade-off between data reliability and emotional granularity. The latter advocates for allowing participants' freedom in their emotional description. It displays a temporal dimension as well: different movements from a symphony will probably convey different emotions.

Given that appraisal theory (emotions are caused by an appraisal of a stimulus) and constructivism (emotion concepts are cognitively and socially constructed) accept the use of self-reports for emotion description (see [2, 15, 16]), we present an updated conceptualization framework of three levels:

- *Dimensional "core affect"*: the core neurophysiological states of simple and non-reflective emotions are arousal and valence, argued to be the basic features of human emotional experience [17] and referring mainly to processes involved in perception [15]. These low-level mapping mechanisms between affect and sound include orientation and embodiment. The former refer to hardwired responses to sound (i.e., brain stem reflex), expectation (i.e., violation of musical structures), and the adjustment of internal oscillators to music (i.e., entrainment). The latter involve the reactivation of past motor and sensory mechanisms – where the body plays a central role in the interaction with music and the environment.

- *Perceived "basic emotions"*: although the number and labels of adjectives for basic emotions are still subject to debate, discrete categories of emotion – such as happiness, anger, sadness – are typically used to annotate perceived emotions [18]. Contextual modifiers begin to play a key role in the diversity of emotion responses: differences in the music itself (e.g., style and lyrics), socio-cultural conventions (e.g., functional uses of music), and individual differences (e.g., listening mood, musical preferences, personality traits, and musical expertise).

- *Induced "complex emotions"*: The domain-specific Geneva Music Scale (GEMS) model has been extensively used to characterize induced emotion in 9 dimensions and has been translated to different languages [19]. In this top layer of abstraction, high-level mechanisms of memory and appraisal are responsible for emotional evaluation: aesthetic judgments, familiarity, episodic memory, identity confirmation, and language differences will now give way to maximum response diversity. It is likely that the descriptions of emotions vary significantly across languages – emotion adjectives like *saudade* or *schadenfreude* do not have straightforward translations and are unlikely to be used homogeneously.

Warrenburg [16] has delivered a thorough comparison on recent musical and psychological emotion theories – a must-read for any beginner to the field since it reveals the diversity (and elusive consensus) of cognitive perspectives on emotion. Extensive research from music psychology has studied how musical properties relate to particular emotions in music [2, 12]. For example: *happiness* relates to fast tempo, bright timbre, and sharp duration contrasts, while *sadness* is linked to slow tempo, legato articulation,

and dull timbre. Although cross-cultural research has shown that some musical features (e.g., tempo and dynamics) may operate relatively similar across cultures with respect to core affects and some basic emotions [18], it is also clear that there are cultural and stylistic idiosyncrasies in the way emotions are expressed even in Western music. When considered in a broader context of cultures and music traditions, there is a need to contextualize the musical features to the cultural conventions of that tradition and attempt to work without the implicit sampling bias of Western music and annotators [20]. Most of the music and emotion research have been carried out by WEIRD participants and researchers (from Western, Educated, Industrialized, Rich, and Democratic backgrounds), which has implications on the generalization of results to a wider audience of listeners.

2.2 Dataset creation – subjective annotation gathering

Subjective listening and rating tasks are the most used strategy to collect, from one or several listeners, music emotion annotations that define the "ground truth" needed – serving as the *output* of machine learning models. Annotators usually listen to a short excerpt of music (around 30 seconds long) and give emotional judgments about the music, based on the taxonomy model selected by the researchers. Typically, annotators are WEIRD music experts (i.e., musicologists or music producers), music enthusiasts (i.e., with no particular experience or knowledge from music), or crowdsourcing platforms. An argument to use music expert annotators for *perceived emotions* is that the response diversity should be diminished, which in turn improves data reliability [3]. Nonetheless, given that the final users of MER systems are likely to be music enthusiasts, collecting annotations from them is also necessary. When several annotators are involved, the resulting "ground truth" is typically the mean/median of all ratings or the adjectives with highest agreement – a caveat regarding this practice is that it works against the proposed personalized and context-sensitive approach. Although more sophisticated practices that consider inter-rater agreement to fuse annotations are common in other fields of affective computing [21], MER datasets typically average the annotations to "ground truths", which unintentionally create the mirage of potential universality to newcomer researchers (see section 3.2).

Namely, a computational algorithm that displays perfect prediction accuracy is correctly predicting the average annotation and not the perception of an individual user.

In this direction, the subjectivity issue has been tackled in the past by processing independent annotations from individual users (personalized MER) or groups of similar users (groupwise MER) [3]. For example, instead of obtaining a "ground truth" with the average of all annotations, a groupwise annotation can be obtained by selecting users with particular characteristics. Several factors can be used to group annotations such as demographics, academic background, language, music experience, and personality. These studies have shown that personalized models significantly improve performance over models trained with an average rating, while groupwise models do not [3] (see section 3 for future directions on annotation gathering and processing schemes). The described process reflects the complexity of obtaining the "ground truth" to MER models, since each annotation produces a heavy cognitive burden to the listeners – particularly for time-varying annotations, as benchmarked by Aljanaki et al. [22]. Gamification strategies and social media tags have been exploited to reduce the cognitive load on annotators, while crowdsourcing platforms like MTurk and Prolific have been used to reduce the difficulty of recruiting participants. The difficulty of annotation gathering also increases due to the aforementioned confusion between the concepts of *perceived* and *induced* emotions [23] – obscuring the outcome of obtaining annotations from social media tags and highlighting the importance of clarifying this distinction to annotators.

We offer the reader a detailed overview of music and emotion datasets in a complementary website (also containing extended bibliography) - outlining the common limitations of dataset construction in MER[1]: type and availability of the data, constrained amounts of annotated music clips, use of diverse taxonomies, lack of common annotation strategies, and few open datasets due to copyright laws. Nonetheless, immense effort has been made by MER researchers to produce datasets that account for context and multi-modal information: lyrics, web-mined semantic information, MIDI transcriptions, and physiological signals.

2.3 Feature extraction – perspectives from signal processing

Theory from section 2.1 has paved the way for signal processing studies in MER that attempt to extract meaningful descriptors for emotion in music – such musical properties can be represented by knowledge-driven features and serve as the *input* to machine

[1] `https://github.com/juansgomez87/datasets_emotion`

learning models. They are considered to be knowledge-driven since they rely on musical theory and psychoacoustics to extract meaningful information from the audio signal. This is typically referred to as "narrowing the semantic gap" between low-level acoustic properties and high-level musical concepts. Drawing from music-theoretic elements, Panda et al. [24] recently reviewed several emotionally-relevant acoustic features. For example, melody relates to fundamental frequency f_0, pitch salience, and register distribution; rhythm relates to note onsets, metrical structure, and note durations; dynamics relate to sound level, loudness, root mean square energy, and note intensity; timbre relates to spectral centroid, mel-spectrum sceptral coefficients, and spectral kurtosis. Following, we describe five tools commonly used for the extraction of features for MER – revealing the diversity of research on musically-related emotions: (1) *MIRToolbox*[2]: a `Matlab` toolbox with acoustic features and statistical descriptors widely used in music cognition research; (2) *OpenSMILE*[3]: this `C++` toolbox offers the IS13 ComParE set (among several descriptors) – a benchmark from speech emotion recognition and inherited to MER; (3) *Essentia*[4]: a `C++` library with a set of acoustic features along with predictions of pretrained classifiers – genre or danceability predictions can be used as context; (4) *PsySound*[5]: a `Matlab` toolbox for the analysis of audio recordings, based on psychoacoustical algorithms; and (5) *Librosa*[6]: a `Python` package for music and audio analysis that has become widely used in recent years due to the rise of deep learning.

As well as acoustic features that describe the physical signal properties, semantic and contextual information are needed to model the expected predictions of computational algorithms (see Figure 1). It is important to highlight that *perceived emotions* are often influenced by extra-musical features, such as lyrics, cultural meanings of the artist, or genre. However, when the objective is to predict *induced emotions*, the role of the musical features is diminished as the impact of episodic memories, functional uses, and listening context increases. Furthermore, contextual information is culture-specific (e.g., tonality and mode are not directly mappable to non-Western music cultures) and multimodal (e.g., gender, age, personality traits, physiological signals, musical preference, familiarity, listening mood, musical expertise, and functional uses of music). MER

[2] `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`

[3] `https://www.audeering.com/opensmile/`

[4] `https://essentia.upf.edu/`

[5] `http://www.psysound.org/`

[6] `https://librosa.org/`

needs to identify and incorporate this relevant extra-musical information.

The relationship between the design and quantification of emotionally-relevant, perceptually plausible acoustic features and contextual descriptors is still an open research topic. Feature engineering is crucial to give way to interdisciplinary endeavors between music cognition and signal processing.

2.4 Evaluation – from feature design to data-driven methodologies

The predominant approaches for emotion taxonomy define the annotation practices to MER dataset creation. In turn, these annotations characterize the algorithms implemented to predict these categories or values. The full dataset (containing features/*input* and annotations/*output*) is split into training and testing sets. Depending on the annotation approach, two prediction problems are considered: (1) in the case of categorical/discrete approaches, classification-based systems are used, in which the algorithms are trained as classifiers; (2) when considering a dimensional/continuous approach, regression-based models are trained to predict values on the dimensional space. The comparison between the prediction outcomes of the algorithm and the original annotations from the test set allow to evaluate the performance of the algorithm. MER models are evaluated with different performance metrics: (1) classification systems report typically accuracy, precision, recall, and F-scores; and (2) regression systems report root-mean-squared error, Pearson's correlation coefficient (ρ), and coefficient of determination (R^2). Yang and Chen [3] offer a comprehensive collection of these classical machine learning methodologies for the MER task. In addition, MER has been collectively evaluated since 2007 in the Music Information Retrieval Evaluation eXchange (MIREX)[7] Audio Mood Classification task – a benchmark strategy to unify evaluation practices [25]. Since 2013, the Multimedia Evaluation Benchmark (MediaEval) [8] has produced several open datasets.

Computational methodologies have displayed a recent paradigm shift: instead of developing knowledge-based emotional features as seen in section 2.3, researchers have begun to use lower-level representations to automatically learn relevant features using data-driven methodologies [26]. In this way, the need to create carefully handcrafted features has been deemed no longer necessary, since deep learning approaches use a

[7] `https://www.music-ir.org/mirex/wiki`

[8] `https://multimediaeval.github.io/`

backpropagation algorithm able to recognize patterns from low-level representations such as spectrograms or raw audio waveforms. This paradigm shift produces two results with impact on MER: (1) research is targeted to musical content processing and tuning of hyper-parameters, while ultimately discarding the most critical contextual data that describe diversity of annotators [4]; (2) generalization to unseen data has been shown to be a major challenge in data-driven models [25].

Finally, two machine learning approaches show promise for the evaluation and improvement of MER models: ensemble learning [27, 28] and active learning [29, 30]. Studies have used ensemble learning – the combination of predictions from multiple machine learning models – to produce expert predictions of different models [23]. According to Panda et al. [24], the predictions from several machine learning models (i.e., genre, danceability) could be used as input features to MER models in order to improve performance. In contrast, user-centric methodologies allow machine learning algorithms to improve performance by using feedback information from the user. For example, active learning minimizes the annotation cost by cleverly choosing unlabeled data instances, such that machine learning algorithms perform better with less training. Active learning can also be used to continuously train a MER model to the annotations of a particular user, leading to personalized models [29] and handling perception uncertainty [30].

3 Future MER and Challenges

Figure 2 reflects the proposed conceptualization framework for MER systems in order to account for each building block discussed in previous sections, while offering a user-centric perspective for MER. Placing the user at the center of the MER system permits to pinpoint where personalized and context-sensitive information exists and should be emphasized: allowing for response diversity in the steps for taxonomy definition and dataset creation, selecting relevant music according to the background of the listener, discovering interpretable/meaningful features for machine learning models, enabling an evaluation feedback on the MER model for evaluation and improvement, and guiding with overall ethical principles with respect to the application scenarios. Moreover, we stress a link between the final evaluation step and the initial taxonomy definition, in order to refine concepts. This allows to highlight where robust methodological standards can help address current challenges, which we elaborate in the following.

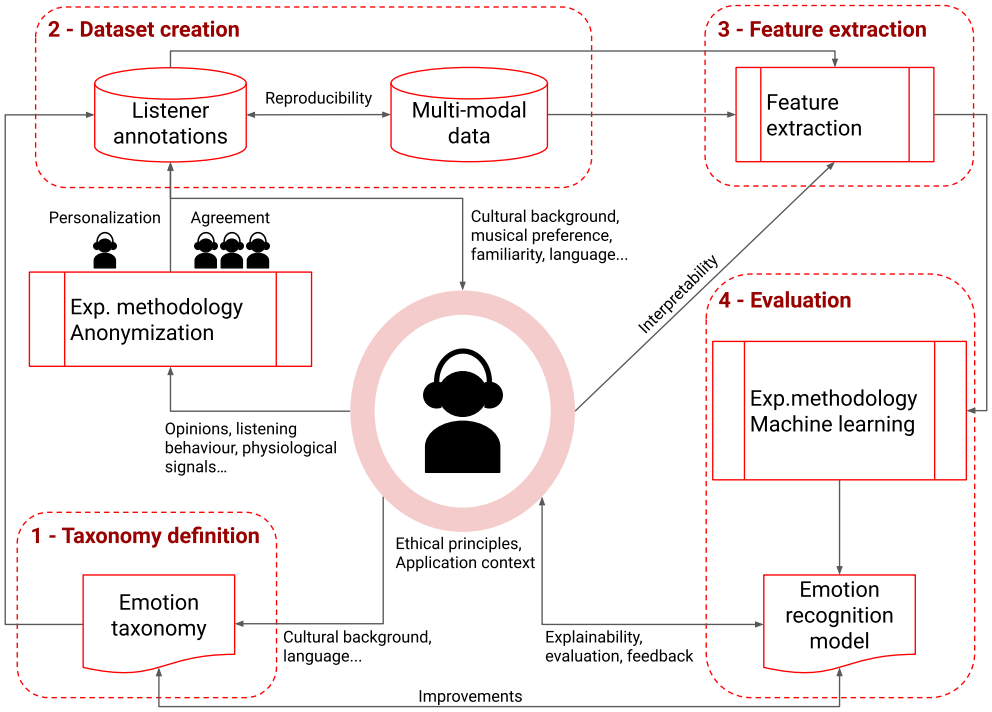

Figure 2: Proposed Music Emotion Recognition systems workflow.

3.1 Data and reproducibility

The availability and development of representative datasets with reproducible methodologies have become fundamental to music and emotion research.

Methodologies. The inclusion of open code and supplementary material to published papers is an increasingly common practice in the MIR community. Nonetheless, deep learning researchers mostly publish trained models, intervening with reproducibility in the cases when the model should be trained on new datasets. Full reproducibility is important for the development of MER systems since it allows for criticism and improvement, as any field that relies on machine learning.

Open data. Since current trends in deep learning require large-scale datasets for training and testing, the availability and development of high quality datasets have become key factors for algorithm accuracy and design. Open datasets have been traditionally limited by copyright laws – interfering with the reproducibility of MER models – since researchers are deterred from sharing raw audio material needed for training and evaluation. Instead of sharing original audio signals, the research community has shared open metadata from the datasets and pre-computed audio features, which have some limitations (e.g., end-to-end learning algorithms). Although the problem of open data

will continue to exist for music research (and alleviated by benchmarking initiatives), novel studies of music editing and music generation could help mitigate this issue. A recent example is EmoteControl[9], an interactive system to control emotional expression of music in real-time. In the future, music generation algorithms with control over musical properties that relate to emotion could produce realistic music creations that convey different emotions, which could be openly shared and useful for model validation.

User data. As pointed out previously, user data (both *user properties* and *user context*) is both limited and needed in order to produce personalized and context-sensitive applications. Streaming services – with the largest amount of data regarding user listening habits – could release the data anonymously to mitigate the issue of data scarcity and benefit research development in this domain. In this direction, the music enthusiast use case from the TROMPA EU project has integrated citizen science strategies to engage with audiences and generate large-scale music emotion annotations [31]. It offers educational material relating music and emotion, while collecting information regarding demographics, language, preference, and familiarity from each annotator. The publication of anonymized user data is highly desirable to enhance reproducibility when MER models begin incorporating personal and contextual data. However, ethical concerns appear from the opposition of user data and privacy – a trade-off between the accuracy in personalization and emotion profiling (see section 3.5).

3.2 Subjectivity and agreement

Quoting Barrett [17] – *variation is the norm* – when describing emotions. This should be taken into account for the design of annotation procedures for MER. Drawing fundamental aspects from music psychology is essential to improve the quality and reliability of subjective annotations. Thus, averaging continuous or Likert-like scale responses to obtain a statistical mean of emotional judgments from a population is a simplification which is often overlooked by different studies of emotional analysis.

Inter-rater agreement. Schedl et al. [6] have described overall low inter-rater agreement of emotion annotations for Beethoven's Eroica symphony, as defined by Krippendorff's α coefficient. Cronbach's α coefficient for internal consistency, also used to assess agreement, relies on correlation coefficients. It has been argued that this coef-

[9] `https://github.com/annaliesemg/EmoteControl2019`

ficient is not suitable to assess data reliability since it favors larger sample sizes [32]. While reliability is critical for the reproducibility of empirical experiments and has a direct impact on algorithmic performance of MER models [33], we suggest that dataset creators should report, analyze, and benefit from inter-rater agreement – the elusive "ground truth" may become an opportunity when "embracing subjectivity" [21, 30].

Annotation schemes. We advocate for the creation of datasets with diverse emotion representations depending on the particular needs: core affects, basic emotions, and complex emotions (as seen in section 2.1). Our recommendation is to collect annotations with both forced-choice categories and free text descriptions in native language – resulting in multi-labeled and language-specific annotations. This will allow to simultaneously capture broad core affects and subtle personal- and cultural-specific variability, while substantially enriching datasets for response diversity. We stress the need of clarifying the concepts of perceived and induced emotion to annotators in order to improve the understanding of the annotation task. The collection of user data (e.g., familiarity, musical expertise, preference, cultural background) and multiple annotations per music excerpt is also desirable, since it allows to perform crucial analyses of response diversity.

Personalization. We highlight the importance of balancing the subjectivity of personal- and cultural-specific conceptualization of perceived and induced emotions. For more than a decade, Yang and Chen [3] have already proposed that the inherent subjectivity of the MER task could be tackled with the use of personalized models, as mentioned in section 2.2. While recent efforts have been made in this direction [27, 28], more future work is needed.

3.3 Explainability and interpretability

Novel deep learning approaches have been criticized for their "black-box" nature – we foresee a shift towards more understandable models.

Data-driven decisions. A general tendency of methodologies has drifted from a knowledge-driven to a data-driven design: MER systems attempt to rely on the music psychology know-how in order to extract meaningful information from physical signals, but more recently the general approach has shifted towards automatic feature learning from deep learning, as mentioned in section 2.2. Nonetheless, we argue that the explainability and interpretability of models are more critical for evaluating subjective

constructions of emotions than other computational tasks. This is due to the potential impact of MER algorithms on data-driven decisions for emotion regulation applications. Hence, well-known and explainable machine learning models are suggested for future researchers – studies attempted to improve the explainability of predictions by using mid-level features [10], and enhance interpretability through source separation to find the impact of different musical voices on perceived emotion [34].

3.4 Cultural and contextual relevance

As mentioned previously, dataset creation for MER systems follows a WEIRD pattern – cross-cultural research is central in order to produce MER models targeted to user groups with different cultural and socioeconomic backgrounds. Several researchers have pointed out important recommendations regarding cross-cultural research [20]: (1) balancing the trade-off between experimental control and ecological validity towards the latter by conducting field studies with non-Western listeners' own living contexts, (2) employing native cultural experts to aid the selection of musical material for dataset creation, and (3) selecting diverse demographic groups a priori by researchers.

Cross-cultural considerations. In this direction, Hu and Yang [33] have studied the impact of dataset size, annotation reliability, and cultural backgrounds on model performance and generalizability for cross-cultural MER – using Western and Chinese pop music annotated by Western and Chinese listeners. Their results reveal that when the size of the training dataset and the annotation reliability of the testing dataset are controlled, MER models are generalizable between datasets sharing a common cultural background of either music or annotators. While it may seem appealing to create and design general models, emerging cross-cultural studies from music psychology and MER have examined how Western-based music descriptors are not applicable for non-Western music. Cultural and sub-cultural implications should be taken into account since tests for gathering annotations suggest that grounding response diversity is a key consideration.

3.5 Ethical implications for MER applications

The High Level Expert Group of the European Commission proposed seven requirements for artificial intelligence systems to be trustworthy[10], which include human agency and

[10] `https://op.europa.eu/s/pInE`

oversight, technical robustness and safety, privacy, transparency, fairness, societal well-being and accountability. Holzapfel et al. [35] discuss some of these ethical dimensions in MIR technologies. We address as follows the requirements that we think are most relevant in the context of MER systems.

Privacy and data governance. To the date, MER systems do not collect personal information, as most resulting "ground truths" are based on averaging emotion annotations. However, some studies have shown the advantage of gathering personal data such as listening habits or physiological signals [36]. In this context, we should define proper anonymization methodologies to ensure full respect for privacy and data protection while taking advantage of this data to build better and personalized models.

Bias and fairness. We have mentioned in the previous sections the Western cultural bias existing in current annotated datasets and listener models, and thus inherited by MER systems, which might lead to unintended discrimination in those systems (e.g., MER systems are less representative of non-Western music or minorities). In this respect, wider collaboration is needed to ensure that MER systems are adapted to different audiences and cultural backgrounds.

Societal well-being. MER, as well as other affect recognition and user profiling systems, have been considered particularly sensitive as they have strong beneficial applications (e.g., mood regulation) but also harmful ones (e.g., to detect vulnerabilities or induce certain emotions). In this sense, evaluating the impact of MER systems on diversity, learning, and well-being domains of the research field can be critical [11].

4 Conclusions

MER is one of the most challenging and alluring research topics since it draws on the key aspects of uses for music: lifting our mood when we are sad, giving us a particular identity and feeling of belonging, reminding us of happy or better times, giving us a way to express ourselves and understand others. The multiplicity of approaches and views regarding music and emotion are rich in diversity, arriving to different theoretical standpoints (and perhaps lack of consensus). However, this lack of consensus reflects exactly the variation of emotional judgments that we attempt to model using computational systems. Our computational approaches should reflect this variation in a "new era of rising affectivism" and data-driven methodologies: improving personalized and context-

sensitive MER models, unifying annotation practices for data gathering, accounting for non-Western music collections and annotators, and enabling for applications that will have a positive impact for the end users. In short, Henry Wadworth Longfellow's famous quote – music is the universal language of mankind – results in a broad generalization that the field of MER will continue striving to overcome. We keep on learning.

References

[1] Tuomas Eerola and Jonna K. Vuoskoski, "A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2013.

[2] Patrik N. Juslin, *Musical Emotions Explained*, Oxford University Press, 2019.

[3] Yi-Hsuan Yang and Homer H. Chen, *Music Emotion Recognition*, CRC Press, 2011.

[4] Markus Schedl, Arthur Flexer, and Julián Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, pp. 523–539, 2013.

[5] Bob L. Sturm, "Evaluating music emotion recognition: Lessons from music genre recognition?," in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, San Jose, USA, 2013, pp. 1–6.

[6] Markus Schedl et al., "On the Interrelation Between Listener Characteristics and the Perception of Emotions in Classical Orchestra Music," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 507–525, 2018.

[7] Elke B. Lange and Klaus Frieler, "Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features," *Music Perception*, vol. 36, no. 2, pp. 217–242, 2018.

[8] Scott Beveridge and Don Knox, "Popular music and the role of vocal melody in perceived emotion," *Psychology of Music*, vol. 46, no. 3, pp. 411–423, 2018.

[9] Björn W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[10] Shreyan Chowdhury et al., "Towards explainable music emotion recognition: The route via mid-level features," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 237–243.

[11] Xiao Hu, Jing Chen, and Yuhao Wang, "University students' use of music for learning and well-being: A qualitative study and design implications," *Information Processing and Management*, vol. 58, no. 1, pp. 1–14, 2021.

[12] Tuomas Eerola and Jonna K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[13] Paul Ekman, "Are there basic emotions?," *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[14] James A. Russell, "A circumplex model of affect," *Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[15] Tuomas Eerola, "Music and emotion," in *Handbook of Systematic Musicology*, R. Bader and S. Koelsch, Eds., chapter Music and Emotion, pp. 539–556. Springer, 2018.

[16] Lindsay A. Warrenburg, "Comparing musical and psychological emotion theories," *Psychomusicology: Music, Mind, and Brain*, vol. 30, no. 1, pp. 1–19, 2020.

[17] Lisa Feldman Barrett, *How Emotions are Made: The Secret Life of the Brain*, Houghton Mifflin Harcourt, 2017.

[18] Petri Laukka et al., "Universal and culture-specific factors in the recognition and performance of musical affect expressions," *Emotion*, vol. 13, no. 3, pp. 434–449, 2013.

[19] Marcel Zentner, Didier Grandjean, and Klaus R Scherer, "Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[20] Nori Jacoby et al., "Cross-cultural work in music cognition: Challenges, insights, and recommendations," *Music Perception*, vol. 37, no. 3, pp. 185–195, 2020.

[21] Jing Han et al., "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," in *Proceedings of the 25th ACM International Conference on Multimedia (MM)*, 2017, pp. 890–897.

[22] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS One*, pp. 1–22, 2017.

[23] Naresh N. Vempala and Frank A. Russo, "Modeling music emotion judgments using machine learning methods," *Frontiers in Psychology*, vol. 8, 2018.

[24] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, pp. 1–20, 2020.

[25] Markus Schedl, Emilia Gómez, and Julián Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[26] Eric J. Humphrey, Juan P. Bello, and Yann Lecun, "Feature learning and deep architectures: New directions for music informatics," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 461–481, 2013.

[27] Marko Tkalčič et al., *Emotions and Personality in Personalized Services*, Springer, 2016.

[28] Yu-An Chen et al., "Component Tying for Mixture Model Adaptation in Personalization of Music Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1409, 2017.

[29] Dan Su and Pascale Fung, "Personalized music emotion classification via active learning," in *Proceedings of the 2nd International ACM Workshop on Music Infor-*

mation Retrieval with User-Centered and Multimodal Strategies (MIRUM), Nara, Japan, 2012, p. 57–62.

[30] Georgios Rizos and Björn W. Schuller, "Average Jane, Where Art Thou? – Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Marie-Jeanne Lesot et al., Eds. 2020, pp. 42–55, Springer.

[31] Gutiérrez-Páez, Nicolás and others, "Emotion Annotation of Music: A Citizen Science Approach," in *Proceedings of the 27th International Conference on Collaboration Technologies and Social Computing (CollabTech)*, Trier, Germany, 2021.

[32] Eunseong Cho and Seonghoon Kim, "Cronbach's Coefficient Alpha: Well known but poorly understood," *Organizational Research Methods*, vol. 18, no. 2, pp. 207–230, 2015.

[33] Xiao Hu and Yi-Hsuan Yang, "Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.

[34] Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 310–317.

[35] Andre Holzapfel, Bob L. Sturm, and Mark Coeckelbergh, "Ethical dimensions of music information retrieval technology," *Transactions of the International Society for Music Information Retrieval*, vol. 1, pp. 44–55, 2018.

[36] Xiao Hu, Fanjie Li, and Jeremy Ng, "On the relationships between music-induced emotion and physiological signals," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 362–369.

Juan Sebastián Gómez-Cañón holds degrees in Music (B.A.), Electronic Engineering (B.Sc.) from the Universidad de los Andes (Colombia), and Media Technology

(M.Sc.) from the Technische Universität Ilmenau (Germany). He joined the Music Technology Group from the Universitat Pompeu Fabra (Spain) as a Ph.D. student under the supervision of Emilia Gómez. His research interests include music emotion recognition and machine learning.

Estefanía Cano holds degrees in Electronic Engineering (B.Sc.), Music (B.A.), Music Engineering (M.Sc.), and Media Technology (Ph.D.). She worked as a research scientist at the Fraunhofer Institute for Digital Media Technology IDMT (Germany), at the Agency for Science, Technology and Research A*STAR (Singapore), and in AudiosourceRe (Ireland). Her research interests are sound source separation, music information retrieval and computational musicology.

Tuomas Eerola is Professor in Music Cognition in Durham University, UK. He has a Ph.D. degree in musicology (music cognition, 2003) and has worked and led topics related to music perception and emotions and music in various externally funded projects (e.g., Academy of Finland and AHRC). At the Department of Music at Durham University, he has served as the Director of Research (2013-2015) and the Head of Department (2018-2020).

Perfecto Herrera holds a B.Sc. in Psychology and a Ph.D. in ICT. He has been with the Music Technology Group, Universitat Pompeu Fabra, Barcelona, since 1997, playing different roles in research activities and projects. He was the Director of the Department of Sonology, Superior Music School of Catalonia (ESMUC), teaching subjects related to music technology. His research interests are audio and music content processing, human/machine classification, and music perception, cognition and emotion, co-authoring more than 150 research papers.

Xiao Hu is an Associate Professor in the University of Hong Kong. Her research interests include music emotion recognition (MER), user-centric music information retrieval, learning analytics and cultural heritage information. She was a tutorial speaker on music affect recognition (2012) and music information retrieval (2016), a Conference Co-chair (2014) and Program Co-chair (2017, 2018) in the International Society for Music Information Retrieval Conference.

Yi-Hsuan Yang is an Associate Research Fellow at the Research Center for IT Innovation, Academia Sinica. He is also with the Taiwan AI Labs as the Chief Music Scientist. He received his Ph.D. in Communication Engineering from National Taiwan

University (2010). He is an author of the book *Music Emotion Recognition* (2011). He was an Associate Editor for the IEEE Transactions on Affective Computing and Transactions on Multimedia (both 2016-2019).

Emilia Gómez is a researcher at the Joint Research Centre, European Commission and Universitat Pompeu Fabra in Barcelona. She holds degrees in Telecommunication Engineering (Universidad de Sevilla), M.Sc. in Acoustics, Signal Processing and Computer Science applied to Music (IRCAM, Paris), and Ph.D. in Computer Science and Digital Communication (UPF). Her research is within the Music Information Retrieval field. Starting from the music domain, she also studies the impact of algorithms into human behaviour.