Music Emotion Recognition: Intention of Composers-Performers Versus Perception of Musicians, Non-Musicians, and Listening Machines

Luca Turchet, *Member, IEEE*, and Johan Pauwels

***Abstract*—This paper investigates to which extent state of the art machine learning methods are effective in classifying emotions in the context of individual musical instruments, and how their performances compare with musically trained and untrained listeners. To address these questions we created a novel dataset of 391 classical and acoustic guitar excerpts annotated along four emotions (aggressiveness, relaxation, happiness and sadness) with three emotion intensity levels (low, medium, high), according to the intended emotion of 30 professional guitarists acting as both composers and performers. A first experiment investigated listeners' perception involving 8 professional guitarists and 8 non-musicians. Results showed that the emotions intended by a composer-performer are not always well recognized by listeners, and in general not with the same intensity. Listeners' identification accuracy was proportional to the intensity with which an emotion was expressed. Emotions were better recognized by musicians than by listeners without musical background. Such differences between the two groups were found for different intensity levels of the intended emotions. A second experiment investigated machine listening performance based on a transfer learning method. To compare machine and human identification accuracies fairly, we derived a fifth, "ambivalent" category from the machine listening output categories (i.e., excerpts rated with more than one predominant emotion). Results showed that the machine perception of emotions matched or even exceeded musicians' performance for all emotions except "relaxation". The differences between the intended and human-perceived emotions, as well as those due to musical training, suggest that a device or application involving a music emotion recognition system should take into account the characteristics of the users (in particular their musical expertise) as well as their roles (e.g., composers, performers, listeners). For developers this translates into the use of datasets annotated by different categories of annotators, whose role and musical expertise will match the characteristics of the end users. Such results are particularly relevant to the creation of emotionally-aware smart musical instruments.**

***Index Terms*—Affective computing, machine listening, music emotion recognition, music information retrieval, musical expertise.**

Manuscript received June 3, 2021; revised September 19, 2021 and November 21, 2021; accepted December 23, 2021. Date of publication December 28, 2021; date of current version January 8, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Isabel Barbancho. *(Corresponding author: Luca Turchet.)*

Luca Turchet is with the Department of Information Engineering and Computer Science, University of Trento, 38122 Trento, TN, Italy (e-mail: luca.turchet@unitn.it).

Johan Pauwels is with the Dyson School of Design Engineering, Imperial College London, SW7 2BX London, U.K. (e-mail: j.pauwels@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2021.3138709

I. Introduction

During the last two decades, the study of emotions in music has attracted an increasing attention of researchers from different disciplines, including experimental psychology and computer science. Emotions in music have been studied from different perspectives: i) perceived, i.e., the emotions identified by an individual when listening; ii) felt, i.e., the emotional responses an individual experiences in body and mind when listening (these, it is worth noticing, can be distinct from the perceived ones); iii) intended, i.e., the emotions that the performer and/or composer aimed to convey [1].

Researchers in the field of Musical Psychology developed different emotion paradigms (e.g., categorical or dimensional [2]), and investigated the capability of encoding emotions in both composers and performers [3], [4] along with the ability of listeners in identifying emotions in music [5]. In parallel, the Music Information Retrieval research community, leveraging results from Musical Psychology, has focused on the topic of Music Emotion Recognition (MER), which aims at devising systems capable of automatically identifying emotions present in musical signals [6]–[8].

Various research gaps exist today for the study of emotions in both domains including, to the best of authors' knowledge, the following. As far as Musical Psychology is concerned: 1) studies so far investigated either the composer or the performer as the source of the intended emotion, not the figure of the composer-performer, who is simultaneously able to create and also express the emotion to be communicated to a listener; 2) the stimuli involved often lack ecological salience [9], e.g., encompassing synthesized sounds, a score composed only by one composer, simple melodic lines, or film music where de-facto the composers were not explicitly instructed to communicate a specific emotion; 3) the influence of the intensity of an intended emotion on the listener's perception has been largely overlooked; 4) the influence of musical training on the ability of listeners to recognize emotions in music is still unclear.

Regarding the gaps in Music Information Retrieval: 1) the vast majority of MER systems have focused on musical pieces involving multiple instruments, while little is known on the performance accuracy of state of the art methods, such as neural networks, on emotional datasets of individual instruments; 2) MER research typically involves datasets that are not created from scratch with the aim of conveying a specific emotion, but generic datasets that are emotionally annotated by a pool of listeners, often leveraging tags derived from users of online platforms; 3) the annotations made by listeners may be biased by their musical expertise, which therefore can lead to MER systems with intrinsic biases that model the emotion perception by specific categories of individuals; 4) MER systems have so far focused on perceived and/or felt emotion, not yet on the modeling of emotions intended by composers-performers; 5) MER methods have not yet taken into account the varying intensity of the intended and perceived emotions.

In this paper we aim to address such gaps, with the end goal of creating a MER method that can be embedded into smart musical instruments and enable novel application scenarios for them. Smart musical instruments [10] is an emerging class of digital musical instruments, which is envisioned to be aware of the emotions expressed by the performers in order to support various kinds of musical activities through dedicated services (such as the query of music repositories by playing excerpts with a given emotional connotation [11]). Specifically, our research questions are: i) how does musical expertise of listeners modulate the emotion decoding ability?; ii) does the intensity of an intended emotion influence the human and artificial identification accuracy?; iii) how well can MER systems identify human intended emotions in music when variation due to instrumentation is removed?; iv) how effective is transfer learning in neural networks for a single-instrument MER task when the donor corpus contains multiple instruments?

To address such questions we first created a specific dataset of intended emotions using the guitar as individual instrument. This dataset was annotated by level of emotion intensity by the same composers-performers who were asked to create and express such emotions. In particular, we involved the classical and acoustic guitar (which respectively use nylon and metal strings) as well as a pool of thirty professional guitarists. We focused on the guitar because it is one of the most widespread and known instrument worldwide and because it is the instrument mostly investigated in smart musical instruments research [10], [11]. Notably, we focused on the figure of composer-performer because we envision the direct application of the investigated methods into musical devices such as smart musical instruments, with which the player commonly creates and expresses emotionally connotated music, such as improvisations (e.g., for recreational music making, performances, rehearsals). Furthermore, merging the roles of composer and performer removes any possibility of emotional ambiguity between composition and performance, such as when a performer interprets a composition in a way that contradicts the composer's intent.

Secondly, we conducted listening tests with another set of professional guitarists as well as with non-musicians, to assess differences in the emotion perception of the two groups. Subsequently, we compared the human identification accuracy with machine learning algorithms, creating models that predict the emotional intent of a composer-performer. We selected a transfer learning approach because we aimed to utilize the state of the art MER model reported in [12], which is freely available and has bindings that allow it to run on an embedded system such as the Elk Audio OS [13] and thus be easily integrated into a smart musical instrument. Before adopting a transfer learning workflow, we attempted other MER methods, but without achieving good performance. Our hypothesis was that the use of the transfer learning MER model reported in [12] coupled with a relatively small ad hoc guitar dataset involving four emotions (aggressiveness, relaxation, happiness and sadness) would have led to satisfactory recognition accuracy, thus enabling the creation of an emotionally-aware smart guitar. We discuss the achieved results in terms of implications for the MER field and how they can be used to inform the design of musical interfaces based on MER systems.

II. Related Work

A. Psychological Studies

The emotional quality of a musical performance is influenced both by the information represented in the musical score (i.e., the contribution of the composer) and by the expressive actions of the performer who interprets the score. Notably, emotions intended by composers and performers may differ [14]. Indeed, composers and performers have different types of musical attributes under their control. Whereas composers primarily control pitch, harmony, tonality, rhythmic structure, and instrumentation, performers focus on the musical microstructure, which comprises subtle variations in timing, playing technique, articulation, loudness, tempo, and often pitch intonation.

A number of researchers have investigated the relation between emotions expressed by a performer and perceived by a listener. The methodology typically adopted by these kinds of studies is that of asking performers to interpret with different emotional intentions some pre-composed melodies or pieces, whereas listeners are asked to assess the presence of a certain emotion among a pool of emotion categories on a scale (e.g., of 11 points) varying from absent to present, or from minimum to maximum. The study reported in [15] involved various instruments (violin, electric guitar, singing voice, and flute) and six emotions (happiness, sadness, fear, anger, tenderness, and expressiveness). Results showed that listeners were generally successful in decoding the intended emotion. The study has been recently replicated in [5] yielding similar results. Comparable results are also reported in [4] for an analogous experiment addressing ornaments of melodic lines, with the exception of happiness which was less recognized.

On the other hand, only a handful of studies have investigated the relation between emotions intended by a composer and the emotions perceived by the listener. In the study reported [3], five composers were asked to compose melodies with six emotional intentions (joy, sorrow, excitement, dullness, anger and peace), which were rendered in the form of synthesized piano sounds. Listeners were asked to provide judgements relating to the emotional quality of the melodies on 7-point Likert scales, one for each of the six emotions. Results showed that composers were capable of communicating distinct and definable emotional qualities to listeners. The adoption of a similar methodology yielded similar results in the study described in [16]. In a different vein, the authors of [2] assessed listeners' perceived emotions of excerpts of film music (composed to convey an emotional intention). Results showed congruence between listeners' reported emotion and intended emotion in film music, which was higher for highly representative examples of the investigated emotions than the moderately ones.

To date there is no consensus on the effect of musical training on musical emotion decoding abilities. Whereas some studies reported no effect of musical training (e.g. [16]–[19]), other studies found an effect of musical training on musical emotion recognition accuracy (e.g. [5], [20]–[23]). This calls for more research on the musical expertise of the listeners as a possible predictor for emotion decoding abilities in music.

B. *Automatic Music Emotion Recognition*

A significant body of research in both Musical Psychology and Music Information Retrieval has focused on the relations between emotions and specific musical attributes, uncovering various associations. For instance, happiness is frequently related to pieces characterized by major modes, whereas sadness and anger are often associated to minor modes [24]; complex, dissonant harmonies are usually associated with emotions such as excitement, tension or sadness, while simple, consonant harmonies with happiness, pleasantness or relaxation [25]. For a recent review on emotionally-relevant audio features for MER see [8], which covers both low-level (e.g., spectral features), perceptual (e.g., articulation), and high-level semantic features (e.g., genre).

A variety of MER technique have been developed e.g., [25]–[29]. A major driving force behind this research is that emotion is consistently ranked as a desirable criterion to search music by [30]. Typically MER tasks have been approached in two different ways. The first consists of regressing a continuous emotional space such as the Arousal-Valence one [31], and subsequently clustering such space to obtain a specific emotion vocabulary [32]. The second comprises the classification of a given musical excerpt into one or more emotions, thus becoming a multi-label classification problem with a fixed vocabulary [33]. In this paper, we focus on the second approach. As shown by results of existing studies [7], [29] and the Audio Mood Classification task of the 2007-2020 Music Information Retrieval Evaluation eXchange, state-of-the-art solutions for multi-label classifications are still unable to accurately solve simple problems such as the classification of four or five emotion classes.

Various MER datasets with emotion annotations have also been proposed, e.g., [27]–[29]. However, such datasets do not take into account the true nature of the emotions intended by the composers and performers (including the intensity level), nor are they annotated according to the perception of emotions of individuals with varying levels of musical expertise which may impact the actual ground truth for MER systems. In this study we are interested in addressing such limitations by focusing on individual instruments rather than considering a dataset of musical ensembles. Large emotionally annotated datasets specific to individual instruments are currently missing, along with dedicated MER methods for such case. This is a major limitation that hampers the development of emotionally-aware smart musical instruments, an emerging family of musical interfaces envisioned in [10].

III. Dataset Creation

One of the aims of this research was to introduce a new, improved dataset, consisting of unfamiliar, thoroughly tested and validated non-synthetic music excerpts, for the study of music-mediated emotions and MER systems. Moreover, this set of stimuli was conceived not only to include examples of target emotions with strong intensities, but also examples with weak intensities that enable the study of more subtle variations in emotion. Notably, we involved completely novel musical pieces because well-known music examples may be familiar to the performers or the listeners, and the resulting elicited emotions can be closely entwined with extra-musical associations [34].

A. *Participants*

To create the emotional guitar dataset we recruited 30 professional acoustic and/or classical guitar players (all Italian, 2 females, 28 males), aged between 25 and 56 (mean = 38.06, SD = 8.83). They reported to have at least 11 years of active music expertise (mean = 26.4, SD = 8.16) and on average started learning playing music at the age of 11. We selected such musicians because they were both able to compose and perform emotional intentions well. Specifically, we aimed to avoid potential differences in the intended emotions that may arise between the two roles [14].

B. *Procedure*

Each guitar player was asked to compose and record at least 12 short emotional pieces, 3 for each of 4 emotions (aggressiveness, relaxation, happiness, sadness). Each recording was required to have a duration ranging from 20 to 50 seconds and should have been performed in optimal conditions such as in a recording studio or a silent room, using the internal microphone system embedded in the instrument or external microphones. Composers were requested to not apply any effect to the guitar signal, but to use the original sound of the instrument. They were asked to create multiple pieces within the same emotion that were distinct from one to another (this was due to our aim to increase variety in the dataset). No further indication was given. Therefore, composers were left completely free to use their creativity to express the indicated emotions, using various levels of emotional intent (e.g., very happy music or a little aggressive piece), playing technique (e.g., fingers or pick), expressive technique (e.g., glissando, bending, tapping, harmonics), style, gender, harmonic progression, tempo, etc. They were compensated € 50.

Some guitarists recorded for both the acoustic and classical guitar, while others recorded more than the 12 compositions required. This led to a total of 391 recordings, of which 259 for acoustic and 102 for classical guitar. Subsequently, composers were asked to indicate for each piece the level of their emotional intent in expressing that emotion, on a 3-point scale indicating a low, medium, and high intensity. Specifically, regarding sadness the values composers could choose from were "a little sad," "sad," "very sad" (analogous for the other emotions). Notably, this request was made after and not before the recording because we wanted to leave the musicians free to express their emotion

TABLE I
NUMBER OF COMPOSED PIECES IN THE CREATED DATASET CATEGORIZED BY THE COMPOSERS' EMOTIONAL INTENT AND ITS INTENSITY

Intensity	Aggressive	Relaxed	Happy	Sad	Total
Low	18	19	27	18	82
Medium	50	40	40	41	171
High	30	39	30	39	138
Total	98	98	97	98	391

with the intensity that they felt was most appropriate, without imposing a particular level on them. Table I provides a description of the dataset in terms of number of composed pieces categorized by the composers' emotional intent and their intensity.

The emotions happiness, sadness, aggressiveness, and relaxation were chosen for two reasons. First, because they have been investigated in several studies on emotional expression in music [35], and because they cover the four quadrants of the two-dimensional Arousal-Valence space [1]. Secondly, because they have been tested in previous machine listening setups [12], [25] (see Section V).

IV. Dataset Annotation: Composers' Intention and Listeners' Perception

The first set of annotations of the dataset are those resulting from the composers' own evaluations of their emotional intention when composing and recording the pieces, which was performed on a scale of 3 levels. Subsequently, we performed a set of listening tests to annotate the dataset according to the perceived emotion. Such tests were also devised to address our research questions of quantifying the difference between the listeners' judgements of the perceived emotion and the original emotional intent of the composers, as well as how such difference may be modulated by the listeners' musical expertise.

A. Participants

Sixteen participants took part to the listening tests, 8 professional guitar players (1 female, 7 males) aged between 27 and 45 (mean = 36.62, SD = 7.06), and 8 non-musicians (2 females, 6 males), aged between 22 and 45 (mean = 27.75, SD = 8.79). The 8 musicians were not involved in the recording of the dataset (their primary instrument was the acoustic guitar for 5 of them, and the classical guitar for the other 3). The average number of years of active practice of guitar playing was 26 years. The average age when starting to learn guitar was 10 years old. The non-musicians had not had any formal or informal instrumental music training, and did not play any instrument. Participants were compensated € 50. None reported any hearing problem.

B. Procedure

To avoid presenting test subjects with an excessive number of test conditions, we divided the 391 recordings into seven blocks of 49 and one of 48. All stimuli were block-randomized. Listeners were asked to wear headphones and judge to what extent they recognized each of the four emotions in each excerpt. For each emotion, listeners performed a rating on a 7-point scale, where each point was labelled. For instance for happiness, the 7 labels were: "very not-happy," "not-happy," "a little not-happy," "neutral," "a little happy," "happy," "very happy" (for the sake of the analysis these labels were converted in the corresponding numbers between -3 and 3). Analogously for the other three emotions. Listeners were instructed to rate the four emotions independently. Each excerpt was presented only once, but listeners could listen to the excerpts as many times as they wanted before giving their judgement. The first test was preceded by a short familiarization phase consisting of two recordings not provided in the main test which were composed and recorded by the first author. All tests were conducted using webMUSHRA, a web-based listening test framework [36].

C. Analysis and Results

Following the analysis paradigm adopted in other studies involving similar listening tests [16], [22], [23], we derived the strongest emotion attributed to each musical piece by each participant. We calculated, for each participant and for each emotion, the percentage of accurate responses, defined as the highest rating score for a piece corresponding to the emotion intended by the composer. When the highest rating corresponded to the label that matched the intended emotion, a score of 1 was given. When the highest rating did not correspond to the emotion, a score of 0 was given. When equally high ratings were given to more than one label, the response was considered as ambivalent and received a score of 0. For example, given a piece composed with the sadness emotional intent and a rating of Aggressive = -2, Relaxed = 2, Happy = -1, Sad = 3, the response would be counted as correct, whereas Aggressive = -2, Relaxed = 3, Sad = 2, Happy = -1, would be counted as incorrect. On the other hand, Aggressive = -3, Relaxed = 1, Sad = 1, Happy = -2, would be considered as ambivalent.

Table II presents the percentage of accurate categorizations for each emotion and the distribution of inaccurate and ambivalent responses for all participants (top), guitarists (middle), and non-musicians (bottom). Fig. 1 illustrates a comparison between the groups of the percentage of accurate categorizations, in total as well as for each emotion. The intended emotions had the highest response percentage for all emotions. As it can be seen on the diagonals of Table II (bold cells), accuracy ranged between 23.71% (happiness in non-musicians) and 64.79% (aggressiveness in guitarists). Nevertheless, ambivalent answers had a high percentage across all conditions and for both groups.

Correct categorizations were analyzed using an ANOVA with chi-square distribution performed on a model fitted with a binomial logistic regression, which had factors emotion and musical expertise, and subject as a random effect. Statistically significant main effects were found for emotion ($\chi^2(3) = 204.15$, $p < 0.001$), musical expertise ($\chi^2(1) = 9.11$, $p < 0.01$), and for the interaction ($\chi^2(3) = 69.39$, $p < 0.001$). Post hoc tests were performed on the fitted model using pairwise comparisons adjusted with the Tukey correction. Regarding the factor emotion, aggressiveness was found to have a higher number of correct answers compared to happiness, relaxation, and sadness (all $p < 0.001$).

TABLE II
MEAN PERCENTAGE AND STANDARD DEVIATION OF THE LABEL THAT RECEIVED THE HIGHEST RATING BY ALL THE LISTENERS (TOP), GUITARISTS (MIDDLE), AND NON-MUSICIANS (BOTTOM), AS A FUNCTION OF THE INTENDED EMOTION OF THE COMPOSER. BOLD INDICATES THE MATCH BETWEEN RESPONSES AND INTENDED EMOTIONS. AMBIVALENT RESPONSES CORRESPOND TO HIGHEST RATINGS GIVEN TO MORE THAN ONE LABEL

	% Responses of all participants				
Intention	Aggressive	Relaxed	Happy	Sad	Ambivalent
Aggressive	**57.07 (±17.94)**	3.57 (±4.14)	5.93 (±4.77)	4.84 (±3.75)	28.57 (±12.29)
Relaxed	4.08 (±3.42)	**35.33 (±14.38)**	7.78 (±5.41)	11.73 (±6.45)	41.07 (±14.57)
Happy	6.7 (±7.46)	14.23 (±12.23)	**37.82 (±19.91)**	3.99 (±3.53)	37.24 (±11.68)
Sad	4.71 (±5.51)	17.53 (±12.14)	2.23 (±1.81)	**36.67 (±14.19)**	38.83 (±11.8)
	% Responses of professional guitarists				
Intention	Aggressive	Relaxed	Happy	Sad	Ambivalent
Aggressive	**64.79 (±12.69)**	1.4 (±1.24)	3.57 (±2.55)	5.35 (±3.33)	24.87 (±10.33)
Relaxed	4.46 (±3.81)	**35.2 (±15.02)**	4.71 (±2.88)	13.39 (±7.34)	41.96 (±15.47)
Happy	5.54 (±3.45)	8.24 (±7.5)	**51.93 (±14.54)**	2.31 (±1.23)	31.95 (±8.62)
Sad	4.2 (±6.68)	13.01 (±7.98)	1.91 (±1.72)	**42.98 (±12.02)**	37.88 (±10.87)
	% Responses of non-musicians				
Intention	Aggressive	Relaxed	Happy	Sad	Ambivalent
Aggressive	**49.36 (±19.07)**	5.73 (±4.83)	8.29 (±5.28)	4.33 (±4.07)	32.27 (±12.96)
Relaxed	4.46 (±3.81)	**35.45 (±15.02)**	4.71 (±2.88)	13.39 (±7.34)	41.96 (±15.47)
Happy	7.86 (±9.83)	20.23 (±13.08)	**23.71 (±13.54)**	5.67 (±4.21)	42.52 (±11.95)
Sad	5.22 (±3.96)	22.06 (±13.79)	2.55 (±1.83)	**30.35 (±13.36)**	39.79 (±12.6)

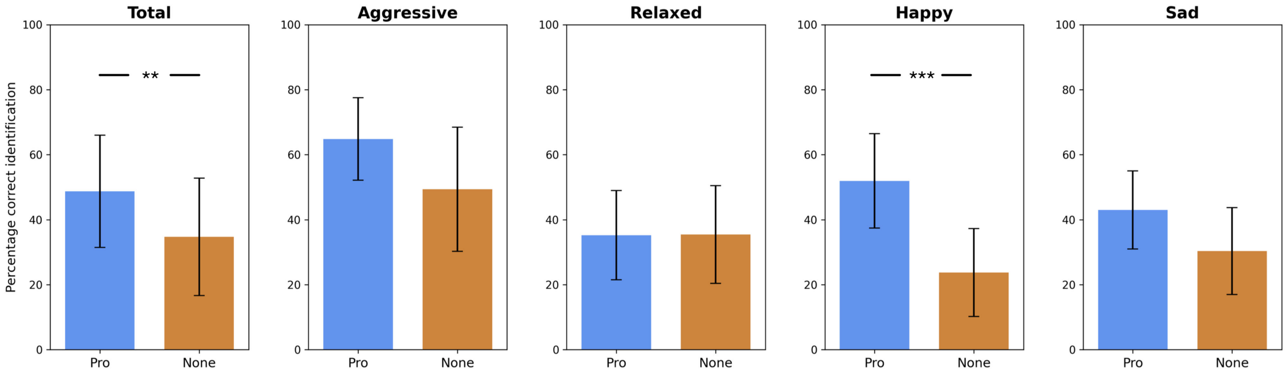


Fig. 1. Mean percentage and standard deviation of the correct identifications as a function of musical expertise level (pro = professional guitarists, none = non-musicians). Legend: ** = $p < 0.01$, *** = $p < 0.001$.

Regarding the interaction term, the number of correct answers of professional guitarists was significantly higher compared to those of non-musicians for happiness ($p < 0.001$).

To assess the differences in participants' identifications due to the emotional intensity, we conducted an analysis on the correct categorizations using an ANOVA with chi-square distribution. This was performed on a model fitted with a binomial logistic regression having factors emotion intensity and musical expertise, and subject as a random effect. Statistically significant main effects were found for emotion intensity ($\chi^2(1) = 42.48$, $p < 0.001$), musical expertise ($\chi^2(1) = 9.55$, $p < 0.01$). Post hoc tests were performed on the fitted model using pairwise comparisons adjusted with the Tukey correction. Regarding the factor emotion intensity, the high intensity was found to have a higher number of correct answers compared to medium intensity ($p < 0.05$) and low intensity ($p < 0.001$), while the medium intensity was found to have a higher number of correct answers compared to low intensity ($p < 0.001$). Regarding the interaction term between musical expertise and emotion intensity, the number of correct answers of professional guitarists was significantly higher compared to those of non-musicians for medium and high intensity (both $p < 0.05$). Results are illustrated in Fig. 2.

We also analyzed the results considering the influence of musical expertise on the difference between the intensity value of the emotion expressed by the composers and the ratings of the listeners, regardless of the identification correctness. This analysis provides a quantification of the error listeners make in identifying the intensity of an emotion originally intended by composers. Results are shown in Fig. 3. We performed an

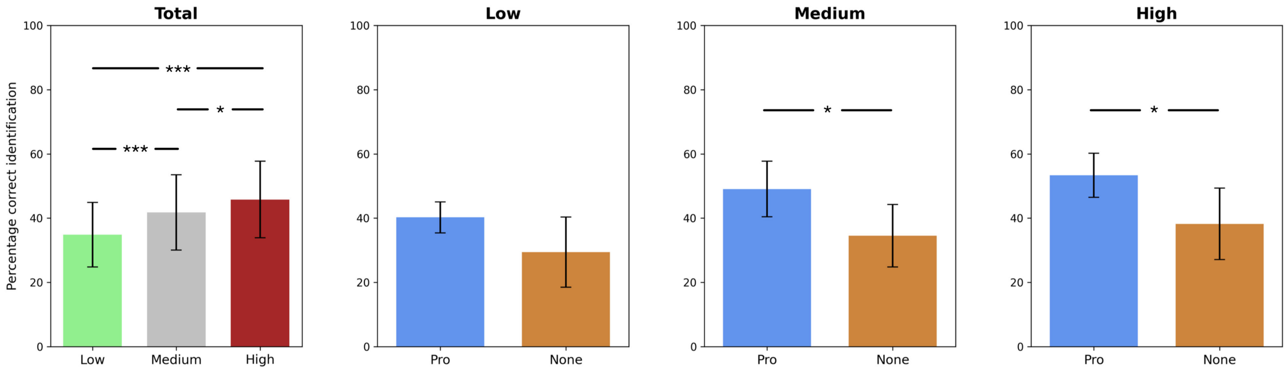


Fig. 2. Mean percentage and standard deviation of the correct identifications as a function of the intensity of the emotional intent of the composer and of the musical expertise level (pro = professional guitarists, none = non-musicians). Legend: * = $p < 0.05$, *** = $p < 0.001$.

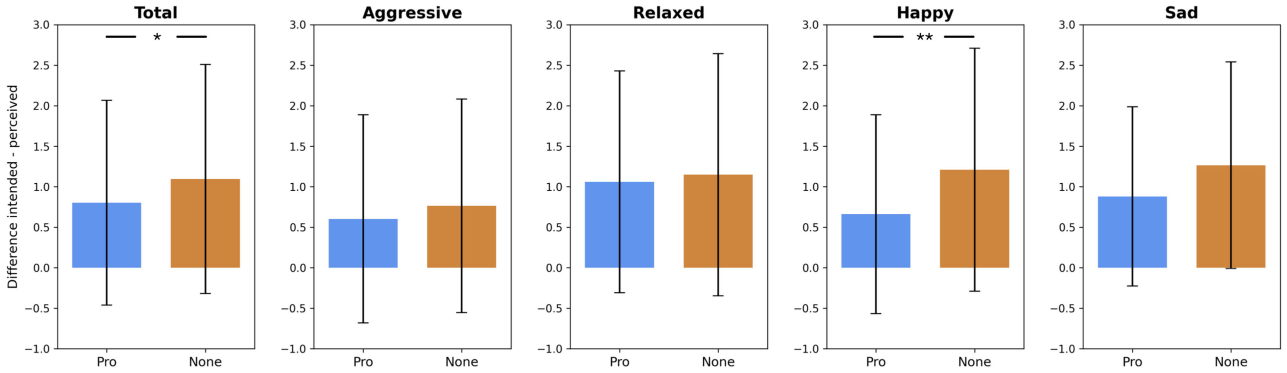


Fig. 3. Mean and standard deviation of the difference between the intensity value of the emotion expressed by the composers and the ratings of the listeners as a function of musical expertise level (pro = professional guitarists, none = non-musicians), regardless of the identification correctness. Legend: * = $p < 0.05$, ** = $p < 0.01$.

ANOVA on a linear mixed effect model having emotion and musical expertise as fixed factors, and subject as a random factor. A significant main effect was found for emotion (F(3,6234) = 33.79, $p < 0.001$), musical expertise (F(1,14) = 4.55, $p < 0.05$), and their interaction (F(3,6234) = 10.21, $p < 0.001$). Post hoc tests were performed on the fitted model using pairwise comparisons adjusted with the Tukey correction. Regarding the factor emotion, aggressiveness was found to have a smaller intended vs. perceived difference compared to happiness, relaxation, and sadness (all $p < 0.001$). Regarding the interaction term, the intended vs. perceived difference of professional guitarists was significantly smaller compared to those of non-musicians for happiness ($p < 0.01$).

D. Discussion

The study investigated the listeners' ability to recognize four distinct emotions (aggressiveness, relaxation, happiness, and sadness) in musical excerpts that were composed to purposely convey these emotions. The musical excerpts were not found to convey effectively the intended emotions, although all emotions were recognized with a better than chance probability by both professional guitarists and non-musicians (in a conventional forced-choice task with five response alternatives, four emotion categories and ambivalent responses, chance-level would be 20%). As shown in Table II (top), the intended emotion was recognized with more than 35% correct for relaxation, happiness, and sadness, while aggressiveness was best recognized (with 57%). Nevertheless, it is worth noticing that ambivalent answers had also a very high percentage across all conditions and for both groups, which is an indication of the uncertainty of listeners in classifying an excerpt as having one dominant emotional component. This is in contrast with the original intention of the composers-performers of expressing unambiguously a determined emotion.

On the other hand, the results clearly indicate an effect of musical expertise on the ability to recognize an emotion intended by a composer. On average, the identification performance of professional guitarists was significantly better than that of non-musicians. However, it should be noted that this does not seem to hold true for all possible emotions. As a matter of fact, the relaxation emotion received percentages of correct identification that were highly similar for professional guitarists and non-musicians.

The results also showed that portrayals with strong emotion intensity yielded higher decoding accuracy than portrayals with weak intensity. Notably, the differences found for the two groups did not depend on the intensity level of the emotion intended by the composers, i.e., they were found for both weak and strong intensities.

V. Automatic Emotion Recognition

A. Experimental Setup

The purpose of this experiment was to investigate how machine learning (ML) models compare to human listeners. Having reliable ML models for emotion recognition would enable the creation of applications based on the indexing of large catalogues of music accordingly, such as for instance the retrieval of emotional music performed via cloud-smart musical instrument interactions as reported in [11].

As noted in the introduction, existing MER systems have so far focused uniquely on processing multi-instrument music. A state of the art system based on deep learning is presented in [12], which has an implementation freely available as part of the Essentia toolkit [37]. The limited generalisation of these multi-instrument models quickly became apparent when running them on our new dataset. All pretrained models available[1] identify the presence of all four emotions in nearly all recordings, resulting in high recall but poor precision.

Models trained specifically on solo guitar are therefore needed. We followed the approach taken by [12] in that we took a pretrained convolutional model created for auto-tagging, named *musicnn* [38], as a donor for transfer learning. We then built a classifier by adding two dense classification layers and an output layer on top of its penultimate layer. The architecture for *musicnn* consists of convolutional and pooling layers that are specifically tuned to capture musical characteristics [39]. Its parameters are set by training on two large music datasets (one set of weights per dataset), MagnaTagATune [40] ($\approx 19\,k$ tracks) and the Million Song Dataset [41] ($\approx 200\,k$ tracks). As our dataset is comparatively small for training a deep learning model, leveraging pretrained models through transfer learning proved to be crucial.

Unlike [12], we trained a single model capable of predicting the most dominant out of the four considered emotions. We used 100 dense nodes with ReLU activation in the penultimate layer and an output layer of four. A softmax function was chosen as activation of the output layer. In comparison, [12] created four different binary classifiers for each emotion, each trained on a separate dataset collected during earlier work on the topic [25]. Our unified dataset makes it possible to consider all emotions at the same time and study their interdependence.

Our choice for a transfer learning workflow also determined the input to the model. Like the original *musicnn* network, our input consists of logarithmically compressed mel-spectrograms consisting of 96 mel bands extracted from audio signals downsampled to 16 kHz in Hann windows of 512 samples with 50% overlap. These mel-spectrograms are then presented to the network in disjoint slices covering 3 s (187 frames).

The network was trained using five-fold cross-validation, where all recordings by the same composer-performers were considered as an indivisible unit when determining the folds. As customary for transfer learning, training was performed in two stages. At first, the weights of the donor network were kept fixed until convergence on the validation set, then the whole model was further updated in a fine-tuning stage.

In order to mimic as closely as possible the listening test, only the emotion category was used as target label, not the intensity level. The reasoning is that we want to teach the machine to recognise the intended emotion in the recordings, but leave it free to assess its intensity; in a similar way as human listeners bring their notion of the four emotions to the experiment, established through earlier exposure to music, and then judge the intensity of individual recordings.

To do so, we use a sparse softmax cross-entropy loss calculated directly from the logits of the output layer. The combination of cross-entropy loss and mutually exclusive intent makes sure that weights in the output layer are only updated when the recording contains their corresponding emotion. This is what we desire, since songs with aggressive intent can independently vary in their level of sadness, for instance. In this sense, we treat the task as a multi-class classification while training, but we are interested in the relation between emotions as well, not just the dominant one. Therefore, the outputs for all emotions are reported individually.

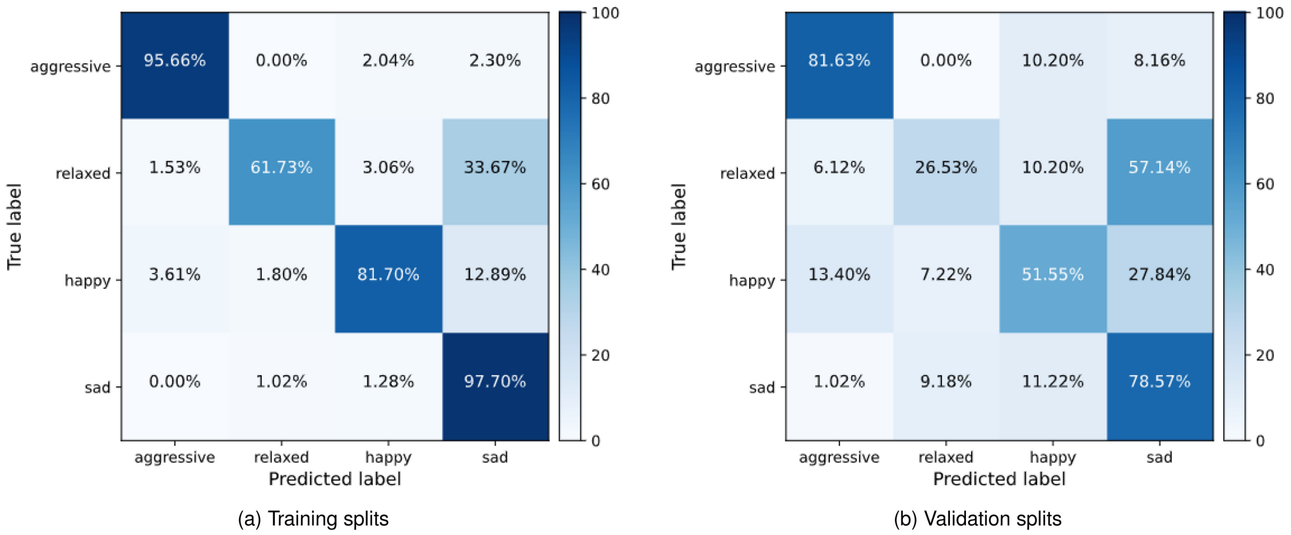
Apart from mimicking the listening test, there are other reasons why we believe the above formulation is the most suitable for this task. The alternative would be to consider a multi-class, multi-label setup, where each of the four emotions can independently take one of seven values. However, doing so would make no use of the ordinal relationship between the different levels of the same emotion. A more pragmatic reason is that it would also require labelling of intent by the composers for all four emotions, including levels of negative intent which would be more challenging, and that it would lead to fewer training examples per class.

B. Results

We used five-fold cross-validation to test which of the MagnaTagATune or the Million Song Dataset was more appropriate as donor for transfer learning, as well as setting some hyperparameters such as batch size and learning rate. The best performing system was obtained with the MagnaTagATune dataset, a batch size of 256 and the Adam optimizer with a learning rate of 0.0001. The mean categorical accuracy including standard deviation of the raw output for this configuration is $77.280 \pm 13.018\%$ for the training splits and $51.756 \pm 3.740\%$ for the validation splits. Since this is the performance on isolated slices of 3 s, not taking into account that the emotion is known to be constant for the whole duration of the recording, we calculated a mean per recording accuracy through soft voting (hard voting was also tested but gave consistently lower results, though still higher than the per slice accuracy). The resulting accuracy is $84.172 \pm 13.044\%$ and $59.269 \pm 5.451\%$ for training and validation, respectively. The confusion matrices for the accuracy after soft voting can be seen in Fig. 4. We can see that aggressiveness is relatively easy to distinguish from other emotions, and that the system has a tendency to overpredict sadness, resulting in a high accuracy for the latter but also high confusion between sad and relaxed. This tendency is already visible in the training data, and gets aggravated in the validation data.

Because the ML model has a continuous output, it has an advantage in that the chance of it returning two emotions with

[1] https://essentia.upf.edu/models/



(a) Training splits

(b) Validation splits

Fig. 4. Aggregate confusion matrices for the optimal network trained with five-fold cross-validation. (a) Training splits (b) Validation splits

TABLE III
PERCENTAGE OF THE LABEL THAT RECEIVED THE HIGHEST RATING BY THE MACHINE LEARNING METHOD AS A FUNCTION OF THE INTENDED EMOTION OF THE COMPOSER. THE RELATIVE DIFFERENCE WITH RESPECT TO FIG. 4B IS GIVEN BY δ. BOLD INDICATES A MATCH BETWEEN PREDICTIONS AND INTENDED EMOTIONS. AMBIVALENT RESPONSES CORRESPOND TO RECORDINGS WHERE THE HIGHEST OUTPUT DIDN'T STAND OUT FROM THE OUTPUTS OF OTHER EMOTIONS

Intention	Aggressive		Relaxed		Happy		Sad		Ambivalent
Aggressive	**75.51**	$\delta: \mathbf{-7.50\%}$	0.00	$\delta: n/a$	7.14	$\delta: -30.03\%$	4.08	$\delta: -50.02\%$	13.27
Relaxed	6.12	$\delta: -0.04\%$	**12.20**	$\delta: \mathbf{-54.02\%}$	6.12	$\delta: -40.02\%$	32.65	$\delta: -42.86\%$	42.86
Happy	10.31	$\delta: -23.07\%$	4.12	$\delta: -42.91\%$	**40.21**	$\delta: \mathbf{-21.99\%}$	17.53	$\delta: -37.02\%$	27.84
Sad	0.00	$\delta: -100.00\%$	3.06	$\delta: -66.68\%$	6.12	$\delta: -45.48\%$	**52.04**	$\delta: \mathbf{-33.77\%}$	38.78

equal intensity level is virtually zero. In contrast, it is clear from Table II that there is much ambivalence in human emotion recognition. In order to make the comparison between ML model and humans more fair, we identify ambivalent responses in the ML output too. To this end, we impose the additional requirement that the output of strongest emotion needs to stand out from the outputs of the other emotions. Since the human annotators had seven levels of intent to choose from, we require the output of the strongest emotion to be at least $\frac{1}{7}$ more than the output of the second highest in order to be considered unambiguous. The outcome of this process on the validation data is displayed in Table III. We can notice the same trends as in the raw output, namely a clear separation of aggressive and an overprediction of sad, particularly affecting relaxed.

The introduction of an ambivalent class understandably reduces the number of correctly recognised recordings for all emotions. However, it also clears up the confusion matrix in the sense that comparatively more of the incorrectly recognised emotions are reassigned to the ambivalent class than correctly recognised emotions. This is apparent from the δs in Table III, which give the relative decrease in recognition accuracy compared to the raw accuracy in Fig. 4(b). We can see that the decrease is stronger for the off-diagonal elements that for the elements on the diagonal, except for the degenerate recognition of relaxed. This indicates that our calculation of emotional ambivalence has the potential to improve the user experience of MER applications by not returning a forced choice when ambivalent emotions are present in a recording, but instead communicating this ambivalence to the user. When an unambiguous emotion is detected, the confidence in this decision will then be higher than when no ambivalence class would be used.

To quantify the significance of the differences in accuracy between emotions, we conducted an ANOVA analysis with chi-square distribution on the output recognised as unambiguously correct. This was performed on a generalized mixed model fitted with a binomial logistic regression having as a factor the composer's intended emotion. The analysis yielded a significant main effect ($\chi^2(3) = 89.75$, $p < 0.001$). Post hoc tests were performed on the fitted model using pairwise comparisons adjusted with the Tukey correction. Aggressiveness was found to have a significantly greater percentage of correct identifications compared to happiness and relaxation (both $p < 0.001$); for happiness the percentage was significantly greater than relaxation ($p < 0.001$); for sadness the percentage was significantly greater than relaxation ($p < 0.001$).

To assess the influence of the levels of intent on the accuracy of the ML model, similarly to the analysis performed on the human annotations, we conducted another ANOVA analysis with chi-square distribution. This was performed on a model fitted with a generalized linear model having as a factor the

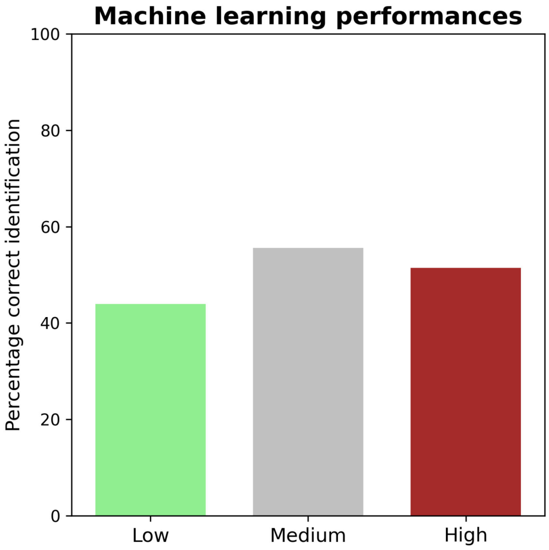


Fig. 5. Percentage of the correct identifications of the ML method as a function of the intensity of the emotional intent of the composer.

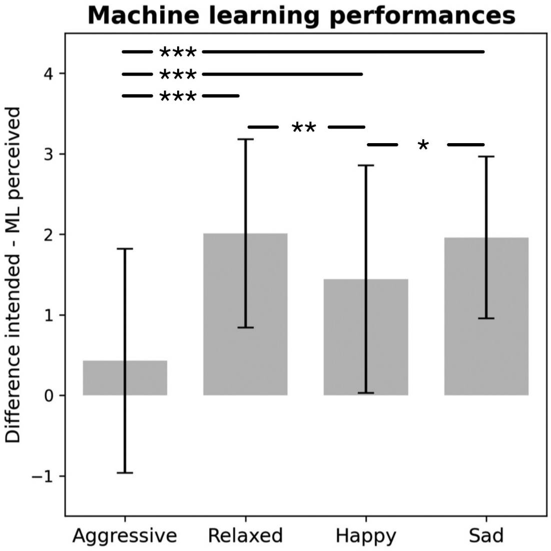


Fig. 6. Mean and standard deviation of the difference between the intensity value of the emotion expressed by the composers and the ratings of the ML method regardless of the identification correctness. Legend: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

composer's intended emotion intensity. No statistical differences were found. The percentages of correct identification are illustrated in Fig. 5.

Furthermore, we performed an analysis by considering the difference between the intensity value of the emotion expressed by the composers and the corresponding raw output values of the ML method, regardless of whether the output of the intended emotion was the maximum over all emotions. This analysis provides a quantification of the error the ML method makes in identifying the intensity of an emotion originally intended by composers. Results are illustrated in Fig. 6. We performed an ANOVA with chi-square distribution on a generalized linear model having emotion as factor. This yielded a significant main effect ($\chi^2(3) = 158.37$, $p < 0.001$). Post hoc tests were performed on the fitted model using pairwise comparisons adjusted with the Tukey correction. Aggressiveness was found to have a significantly smaller intended vs. machine-perceived difference compared to happiness, relaxation, and sadness (all $p < 0.001$), while for happiness the difference was significantly smaller than relaxation ($p < 0.01$) and sadness ($p < 0.05$).

C. Discussion

The emotion recognition performance of the machine learning model heavily depends on the emotion. Whereas aggressiveness and sadness are detected well, the performance on relaxation is quite poor. There are two potential explanations for this behaviour, which are non-exclusive. A first is that the donor model for transfer learning acts as a feature detector. It is possible that the features learnt on the original dataset (in this case MagnaTagATune) are not as suitable for detecting relaxation as they are for other emotions. Although our new guitar dataset is balanced, the comparatively small amount of data is not enough to overcome this limitation during the finetuning stage.

A second reason for the difference in performance could be that the distribution of our training examples is non-optimal. Due to the complex interrelation between emotions, the balanced distribution over composer's intent does not necessarily mean that the overall presence of emotions is balanced. One observation supporting this is that when we try training a model from scratch, without transfer learning, we end up with a degenerate model predicting relaxed roughly 60% and sad 40% of the time, regardless of the emotional intent and for both training and validation splits.

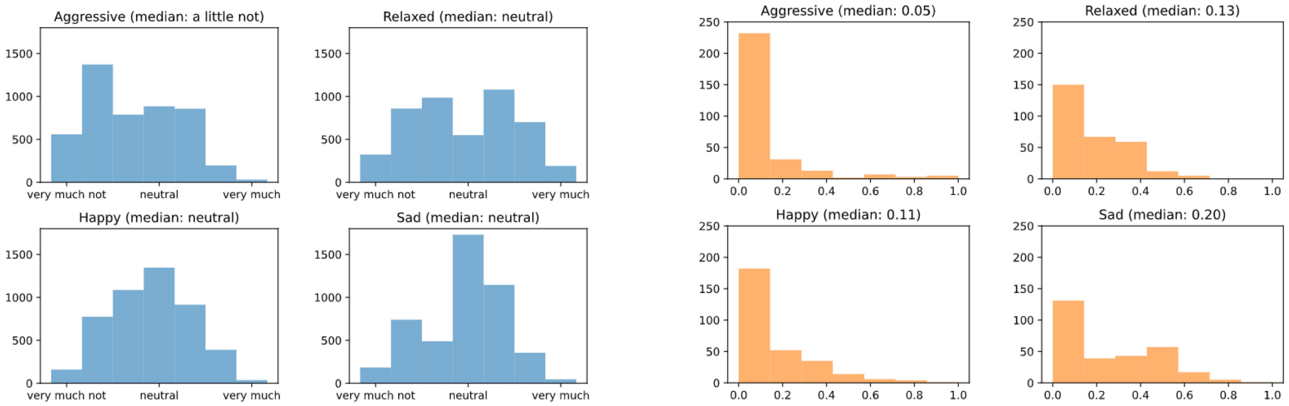
Relaxed and sad also form the most common confusion pair in our final model, although sad is the most predicted emotion there. This reversal of most predicted class is likely due to the difference in training from scratch versus using transfer learning, indicating that both effects are at play and cannot be seen isolated from each other.

That said, if we want to further improve the model based on transfer learning with the *musicnn* model trained on the MagnaTagATune dataset, the best course to take would be the addition of new training examples focusing particularly on sadness and relaxation. The directive to the composer-performers could even be extended to create examples that are "relaxed, but not sad" or similar. Adding extra examples of aggressiveness is unlikely to have to the same benefits per example.

Finally, the level of emotional intent of the composer appears to be of no significance for the ML model. It is not excluded that this level can be learnt if it is explicitly presented as a target to the model, but it does not appear naturally in the output values of the model. Nonetheless, the clearing up of the confusion brought by the introduction of an ambivalent class based on the output demonstrates that the output values provide useful information about the model.

VI. General Discussion

Overall, the emotion recognition performance of professional guitarists was significantly higher than that of non-musicians,



(a) Human ratings

(b) Machine output

Fig. 7. Perception of emotions in recordings not explicitly intended to convey that emotion, for (a) both human ratings and (b) machine output.

suggesting the effectiveness of musical expertise in modulating the ability to recognize an emotion intended by a composer-performer. This finding is consistent with similar studies reported in the literature [22], [23], which however involved the less ecologically-valid dataset presented in [16]. In addition, these results are in line with several previous studies demonstrating that musical expertise leads to brain plasticity and is effective in improving music processing as measured by pitch, timing and timbre discriminations [42]. As the recognition of emotions in music is based on psychoacoustic cues and musical features, better identification accuracies were expected for professional guitarists compared to non-musicians.

Our results also demonstrated that the intensity of the emotion intended by composers-performers had an effect on the identification performances of participants, where significantly higher recognition accuracies were found for emotions composed and expressed with higher intensity than those composed and expressed with lower intensities (see Fig. 2). This result is in line with the findings reported in [2], which involved both moderately and highly representative examples of five discrete emotions, as well as with those of similar non-musical studies conducted on vocal expression of emotions [43]. These findings suggest that the lack of control for emotion intensity may account for some of the inconsistencies in identification accuracies and cue utilization reported in the literature. Moreover, differences due to musical expertise were effective in modulating the identification performances for all the three intensity levels.

Nevertheless, while the identifications of listeners were better than chance, they were not highly accurate, especially for non-musicians. Unintended emotional qualities were judged to be present in the excerpts in varying degrees and a consistent portion of the intended emotions was judged by listeners as ambivalent. This finding differs from others present in the literature investigating emotions conveyed by individual instruments (such as [23] and [22]), which utilized synthesized stimuli varying only the structural details of the composition (e.g., mode, dissonance) and not the performance-related expressive features (e.g., dynamics, attacks) [16]. In contrast, our accuracies are in line with those reported in [5], which employed a similar methodology, although involved only simple melodic lines played by different individual instruments.

As for the performance of a machine learning model, both similarities and differences with respect to human listeners can be identified. Similarly to human listeners' performances, aggressiveness is the easiest emotion to identify. In contrast, the intensity of emotional intent has no relevance for the performance of the ML model. Based on the results in Fig. 4(b), the ML model seems to significantly outperform professional musicians on all emotions except relaxation. However, this is partly due to the model having continuous output, therefore virtually always avoiding ties. The introduction of the concept of ambiguity in Table III gives a more nuanced picture in its comparison with Table II. The amount of ambivalent recordings has a similar range for human and ML ratings, strengthening our belief that our derivation for the latter is sensible. Subtracting these ambivalent recordings from the ML output, the performance on aggressive and sad recordings is still clearly better than human recognition, but happiness gets recognised poorly compared to a professional musician, and is now in line with the general population. However, it should be noted that the relatively high standard deviation of the human ratings makes it hard to conclude anything decisively.

One observation that could explain the relative ease to recognise aggressiveness for humans and machines alike is the difference in perception of an emotion when it is not explicitly intended. In Fig. 7, the ratings and output values of the four emotions are displayed for those recordings that intend to convey another emotion (so the set of recordings differs between emotions). In Fig. 7(a), we see that the perception of all emotions except aggressiveness are symmetric, concentrated around "neutral," meaning that in the absence of explicit intent, these emotions are perceived present or not present to an equal extent. The case of aggressiveness is different though: when not explicitly intended, aggressiveness is perceived more absent than present. The median values of perception confirm this: aggressiveness has a median of "a little not" whereas the other emotions have "neutral" as median value when they are not explicitly intended. Similarly, the output values of the ML model, shown in Fig. 7

b, are much more concentrated at zero for aggressive than for other emotions. Both cases indicate that there are more negative examples of aggressiveness present in the dataset than of other emotions, due to the interrelationships between emotions, and this larger contrast makes it easier to recognise aggressiveness.

The differences between the intended and perceived emotions, as well as those due to musical training, suggest that deciding whether a device or application involving a MER system is appropriate for the task should take into account the characteristics of the users (in particular their musical expertise) as well as their roles (e.g., composers, performers, listeners). For instance, if the system is a smart guitar recognizing the emotions expressed by the player (e.g., for making a query by emotion, similar to [11]) then it is appropriate to train a machine learning model on a dataset of emotions intended by composers-performers. Conversely, the same model would not be appropriate if the system is a music recommendation application for a music streaming service, especially if the listener does not have musical expertise.

It should be noted that the present study involved mostly Italian participants for both roles of composer-performer and listener. However, cultural differences may impact the way in which emotions in music are intended and perceived [44]. Another limitation of our study is the focus on the classical and acoustic guitar. Further research is thus needed to investigate the research questions here addressed involving participants belonging to different cultures along with different individual musical instruments.

VII. Conclusion

The primary objective of the present study was to investigate to which extent state of the art MER methods are effective in modeling emotions in the context of individual musical instruments. Our investigation required us to distinguish between intended and perceived emotion and collect human annotated data for both. These human ratings were then used to conduct a study regarding the effectiveness of communicating emotion from composer-performer to listener. It also served as a point of reference to compare the machine learning model with.

Our results show that the emotion intended by a composer-performer are not always well recognized by listeners, and in general not with the same intensity. The intensity with which an emotion was expressed was proportional to the accuracy of the listeners. In particular, we found that musical expertise affects the perception of emotions in music: emotions were better recognized by musicians rather than listeners with no musical background with respect to the original intention of the composer. Such differences between the two groups were found for different intensities levels of the intended emotions.

No such relation with emotional intensity was observed for the ML model. For three out of four emotions, the machine perception of emotions matched or even exceeded human performance, but relaxation proved to be difficult to learn for the model, already during its training stage. Two possible causes were postulated, namely limitations on the features learnt by the donor model and data imbalances arising from complex interrelations between emotions. The latter can potentially be remedied by adding new training data created according to precise directives. Meanwhile, the output values of the model can be used to identify ambivalent emotions.

In future work we plan to extend the results of the present study by utilizing different types of individual musical instruments as well as involving participants from different cultures. We also plan to create MER models for the listener's perception, potentially on the level of individual users. Further areas to explore include the creation of models based on the specific expertise of the composer-performer, distinguishing beginners and intermediate musicians. Finally, we plan to improve the accuracy of the MER systems by including the latest ML techniques such as attention and compare them with systems based on handcrafted features such as the ones described in [29].

The raw data supporting the conclusions of this manuscript will be made available by the authors to any qualified researcher.

References

[1] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. London, U.K: Oxford Univ. Press, 2001.

[2] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychol. Music*, vol. 39, no. 1, pp. 18–49, 2011.

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

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

[5] J. Akkermans *et al.*, "Decoding emotions in expressive music performances: A multi-lab replication and extension study," *Cogn. Emotion*, vol. 33, no. 6, pp. 1099–1118, 2019.

[6] Y. Kim *et al.*, "Music emotion recognition: A state of the art review," *Proc. Int. Soc. Music Inf. Retrieval Conf.*, vol. 86, pp. 937–952, 2010.

[7] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Syst.*, vol. 24, no. 4, pp. 365–389, 2018.

[8] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.3032373 .

[9] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychol.*, vol. 5, no. 1, pp. 1–29, 1993.

[10] L. Turchet, "Smart musical instruments: Vision, design principles, and future directions," *IEEE Access*, vol. 7, pp. 8944–8963, 2019.

[11] L. Turchet, J. Pauwels, C. Fischione, and G. Fazekas, "Cloud-smart musical instrument interactions: Querying a large music collection with a smart guitar," *ACM Trans. Internet Things*, vol. 1, no. 3, pp. 1–29, 2020.

[12] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in essentia," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 266–270.

[13] L. Turchet and C. Fischione, "Elk audio OS: An open source operating system for the Internet of Musical Things," *ACM Trans. Internet Things*, vol. 2, pp. 1–18, 2021.

[14] L. Quinto and W. Thompson, "Composers and performers have different capacities to manipulate arousal and valence," *Psychomusicol., Music Mind Brain*, vol. 23, no. 3, pp. 137–150, 2013.

[15] A. Gabrielsson and P. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychol. Music*, vol. 24, no. 1, pp. 68–91, 1996.

[16] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon, and B. Bouchard, "Happy, sad, scary and peaceful musical excerpts for research on emotions," *Cogn. Emotion*, vol. 22, no. 4, pp. 720–752, 2008.

[17] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet, "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts," *Cogn. Emotion*, vol. 19, no. 8, pp. 1113–1139, 2005.

[18] P. Juslin, "Emotional communication in music performance: A functionalist perspective and some data," *Music Percep.*, vol. 14, no. 4, pp. 383–418, 1997.

[19] J. Hailstone, R. Omar, S. Henley, C. Frost, M. Kenward, and J. 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.

[20] P. Juslin, "What does music express? basic emotions and beyond," *Front. Psychol.*, vol. 4, 2013, Art. no. 596.

[21] M. Park *et al.*, "Differences between musicians and non-musicians in neuro-affective processing of sadness and fear expressed in music," *Neurosci. Lett.*, vol. 566, pp. 120–124, 2014.

[22] S. L. Castro and C. Lima, "Age and musical expertise influence emotion recognition in music," *Music Percep., An Interdiscipl. J.*, vol. 32, no. 2, pp. 125–142, 2014.

[23] A. Sharp, M. Houde, B. Bacon, and F. Champoux, "Musicians show better auditory and tactile identification of emotions in music," *Front. Psychol.*, vol. 10, 2019, Art. no. 1976.

[24] A. Gabrielsson and E. Lindström, "The Influence of Musical Structure on Emotional Expression," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. London, U.K: Oxford Univ. Press, 2001, pp. 223–248.

[25] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: Design and integration of an automatic content-based annotator," *Multimedia Tools Appl.*, vol. 48, no. 1, pp. 161–184, 2010.

[26] C. Laurier, P. Herrera, M. Mandel, and D. Ellis, "Audio music mood classification using support vector machine," *MIREX Task Audio Mood Classification*, pp. 2–4, 2007.

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

[28] A. Aljanaki, Y. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS one*, vol. 12, no. 3, 2017, Art. no. 173392.

[29] R. Panda, R. Malheiro, and R. Paiva, "Novel audio features for music emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 614–626, Oct./Dec. 2020.

[30] C. Inskip, A. Macfarlane, and P. Rafferty, "Towards the disintermediation of creative music search: Analysing queries to determine important facets," *Int. J. Digit. Libraries*, vol. 12, no. 2, pp. 137–147, Aug. 2012.

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

[32] M. Soleymani, M. Caro, E. Schmidt, C. Sha, and Y. Yang, "1000 songs for emotional analysis of music," in *Proc. ACM Int. Workshop Crowdsourcing Multimedia*, 2013, pp. 1–6.

[33] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards explainable music emotion recognition: The route via mid-level features," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 237–243.

[34] J. Vuoskoski and T. Eerola, "Extramusical information contributes to emotions induced by music," *Psychol. Music*, vol. 43, no. 2, pp. 262–274, 2015.

[35] A. Gabrielsson and P. N. Juslin, "Emotional expression in music," in *Proc. Handbook Affect. Sci.*, R. J. Davidson, H. H. Goldsmith, and K. R.E. Scherer, Eds. Oxford University Press, 2003, pp. 503–534.

[36] M. Schoeffler *et al.*, "webMUSHRA–A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, 2018.

[37] D. Bogdanov *et al.*, "Essentia: An audio analysis library for music information retrieval," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 493–498.

[38] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," in *Proc. Late-Breaking/Demo Session Int. Soc. Music Inf. Retrieval Conf.*, 2019.

[39] J. P. Puig, O. N. Caballero, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 637–644.

[40] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music annotation," in *Proc. 10th Int. Conf. Music Inf. Retrieval*, 2009, pp. 387–392.

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

[42] N. Kraus and B. Chandrasekaran, "Music training for the development of auditory skills," *Nature Rev. Neurosci.*, vol. 11, no. 8, pp. 599–605, 2010.

[43] P. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, no. 4, pp. 381–412, 2001.

[44] H. Argstatter, "Perception of basic emotions in music: Culture-specific or multicultural?," *Psychol. Music*, vol. 44, no. 4, pp. 674–690, 2016.

Luca Turchet (Member, IEEE) received the master's degrees (*summa cum laude*) in computer science from the University of Verona, Verona, Italy, in classical guitar and composition from the Music Conservatory of Verona, Verona, Italy, and in electronic music from the Royal College of Music of Stockholm, Stockholm, Sweden, and the Ph.D. degree in media technology from Aalborg University Copenhagen, Copenhagen, Denmark. He is currently an Assistant Professor with the Department of Information Engineering and Computer Science of University of Trento, Trento, Italy. He is the Co-Founder of Elk and an Associate Editor for the IEEE ACCESS and of the *Journal of the Audio Engineering Society*. His scientific, artistic, and entrepreneurial research has been supported by numerous grants from different funding agencies including the European Commission, European Institute of Innovation and Technology, European Space Agency, Italian Minister of Foreign Affairs, and Danish Research Council.

Johan Pauwels received the Master of Science degrees in electrical/electronics engineering and artificial intelligence from Katholieke Universiteit Leuven, Leuven, Belgium, in 2006 and 2007, respectively, and the Ph.D. degree in automatic harmony recognition from audio from Ghent University, Ghent, Belgium, in 2016. He is currently a Research Fellow with the Audio Experience Design Group, Imperial College London, London, U.K. His main research interests include music information retrieval, signal processing, AI and big data applied to music. He has contributed to multiple national and international research projects in the U.K., Belgium, and France funded by the European Commission and U.K. Research and Innovation.