Examining Emotion Perception Agreement in Live Music Performance

Simin Yang, Courtney N. Reed, Elaine Chew, and Mathieu Barthet

Abstract—Current music emotion recognition (MER) systems rely on emotion data averaged across listeners and over time to infer the emotion expressed by a musical piece, often neglecting time- and listener-dependent factors. These limitations can restrict the efficacy of MER systems and cause misjudgements. We present two exploratory studies on music emotion perception. First, in a live music concert setting, fifteen audience members annotated perceived emotion in the valence-arousal space over time using a mobile application. Analyses of inter-rater reliability yielded widely varying levels of agreement in the perceived emotions. A follow-up lab-based study to uncover the reasons for such variability was conducted, where twenty-one participants annotated their perceived emotions whilst viewing and listening to a video recording of the original performance and offered open-ended explanations. Thematic analysis revealed salient features and interpretations that help describe the cognitive processes underlying music emotion perception. Some of the results confirm known findings of music perception and MER studies. Novel findings highlight the importance of less frequently discussed musical attributes, such as musical structure, performer expression, and stage setting, as perceived across audio and visual modalities. Musicians are found to attribute emotion change to musical harmony, structure, and performance technique more than non-musicians. We suggest that accounting for such listener-informed music features can benefit MER in helping to address variability in emotion perception by providing reasons for listener similarities and idiosyncrasies.

Index Terms—Music and emotion, music perception, inter-rater reliability, individual factors, live performance, music emotion recognition, music information retrieval

1 Introduction

Music, like other forms of art, is subjective and response to music is ultimately up to individual interpretation. Music can both *convey* and *evoke* emotions [1]. Some common approaches used in the investigation of these emotions involve self-reporting [2], through which participants can actively report their own subjective experiences. This may include *perceived emotion*, that which the listener recognises the music is trying to convey, or *induced emotion*, that which is felt by the listener in response to the music [3]. A single musical work can express a range of emotions that vary over time and across individual listeners [4], [5], [6]; thus, self-reporting investigations may use time-based annotation of emotions to help identify detailed, localised emotion "cues" [7], [8], [9], [10].

Previous work using listener annotations has determined that music features such as dynamics, tempo, mode, melodic-harmonic progression and interactions, and sound articulation impact perceived emotion [11], [12]. Continuous-time music emotion recognition (MER) focuses heavily on mapping musical features or low-level correlates to continuous emotion data [13], [14], [15]. Current machine learning approaches may efficiently predict listener perception, but may also face confounding model performance [16], [17], and often fail to address underlying cognitive processes [18], [19]. Although low-level acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), relate to timbre perception [20] and are commonly used in predictive emotion models [13], [21], [22], it is unknown how these features influence perceived emotion and the features do not submit readily to cognitive modelling [23], [24].

In the attempt to develop computational models linking music and associated emotions, the subjective and unique perspective of each individual listener has rarely been taken into account [2], [25], [26]. Music emotion research often requires the assessment of agreement among listeners; however, agreement in music emotion ratings from multiple listeners is usually limited [16], [27], [28]. Variance between listeners can be caused by numerous factors, including the inherent subjectivity of individual perception, participants' limited understanding of emotion taxonomies, ill-defined rubrics used to rate emotion and insufficient rating training, and the lack of controls on data collection when using online or crowd-sourcing platforms. MER normally utilises *an average* or *majority emotion response* as a target for explanatory or predictive models, or simply discards inconsistent ratings from further investigation. This is a reductive way of examining the problem; we must first understand the reliability of emotion annotations, as findings of disagreement between raters are still useful and may indicate that emotion perception differs among individuals more than previously thought [29]. This is evident in the MediaEval Database for Emotional Analysis in Music (DEAM) [16]: the limited consistency in annotator ratings poses a reliability issue when using averaged emotion annotations as "ground truth" for the creation of a generalised model [30]. This has driven analyses instead towards investigation of the differences between annotators. In emotion modelling, the divergence between participant annotations from this generalisation produces a natural upper bound for computational approaches and

- *S. Yang, C. N. Reed, and M. Barthet are with the Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom.*
- *E. Chew is with the CNRS-UMR9912/STMS, Institute for Research and Coordination in Acoustics/Music (IRCAM), Paris, France.*

Manuscript received Month XX, 2020; revised Month XX, 2021.

creates a serious bottleneck in MER system performance [31]. Models that would go beyond what humans agree upon perhaps lead to a systematic misrepresentation of how emotion perception occurs in an empirical setting [29]. The present work was thus driven by the research questions: (1) *"Does the agreement between listeners on perceived emotion vary over the course of a musical piece?"*, (2) *"What are the connections between the emotions perceived and the observed semantic features*[1]*?* (3) *"What musical or individual factors contribute to specific perceived emotions and emotion change?"*

In an initial study, time-based valence-arousal ratings were collected during a *live* music performance (Live study). In a secondary study, these emotion ratings were explored through open-ended feedback from participants in a *controlled lab* setting (Lab study). Through joint thematic analysis [32] of participants' feedback built upon previous findings [33], we have identified *seven key themes* influencing emotion annotations. The analysis highlights the importance of features such as instrumentation, musical structure, expressive embellishments, and music communication as being more closely aligned with underlying cognitive processes. We thus propose a more comprehensive focus in music emotion modelling to include these listener-informed features. We believe attention to underlying semantic themes will address emotional inconsistencies and redirect the focus of MER systems to the listener experience. Through the Lab study, we also investigate how listeners' music backgrounds influence the cognitive processes underlying music emotion perception. We provide a comprehensive summary of the connections between listener-based features and related music information retrieval (MIR) features by listing existing extraction tools and related computational works. Finally, we explore how the setting (live vs. lab) can potentially influence music emotion perception over time regarding agreement levels and rating frequency.

2 Live Study: Time-Based Audience Emotion

An initial Live study was conducted to explore agreement in time-varying emotion ratings across audience members. The listeners annotated their emotions in real-time with the use of a web-based mobile application during a live music performance.

2.1 Materials & Apparatus

2.1.1 Live Music Performance Context

Live music performance conducted in an ecological setting may yield stronger emotion cues and enhance listener experience, compared to recorded performances. This can be due to the presentation of information found in the day-to-day experiences of emotion, particularly in the performer's body language, movement, and facial expression [34], [35]. The setting of a performance and the behaviour of the audience also give context to the music—different venues and musical genres have individual cultures and impose distinct expectations on concert goers, which may elicit different musical responses [36], [37]. The use of live performance thus provides a shared emotional listening context for the audience.

1. Semantic features refer to the meaning expressed by music that can be characterised in linguistic terms.

2.1.2 Music and Setting

The music selected for this study was Arno Babajanian's (1921 - 1983) *Piano Trio in F# minor* (1952) performed by Hilary Sturt (violin), Ian Pressland (cello), and Elaine Chew (piano), with simultaneous visualisation of spectral art by Alessia Milo. The piece was performed twice at the 2015 Inside Out Festival on 22 October at Queen Mary University of London (QMUL). Audio and video were recorded by Milo, and the first performance was chosen for this study[2].

The approximately 23-minute piece presents three movements with widely contrasting tempos (**Table 1**) and is not well known to general audiences, thus likely to avoid familiarity bias. The piece is still firmly within the Western classical tradition. This allows us to relate the present research to the majority of related MER research [38]; however, the perception of this musical style may not be relevant to other genres, as addressed in Section 5.2.

Movement	Duration (Min:Sec)	Tempo Marking	Tempo Characteristics
1	10:14	*Largo*	slow
		Allegro expressivo	fast, bright, expressive
		Maestoso	majestic
2	6:15	*Andante*	walking pace, steady
3	7:20	*Allegro vivace*	rapid, lively

Tbl. 1: The three movements of Babajanian's *Piano Trio in F# minor* with performed duration, composed tempo markings, and respective characteristics.

2.2 Annotation Interface

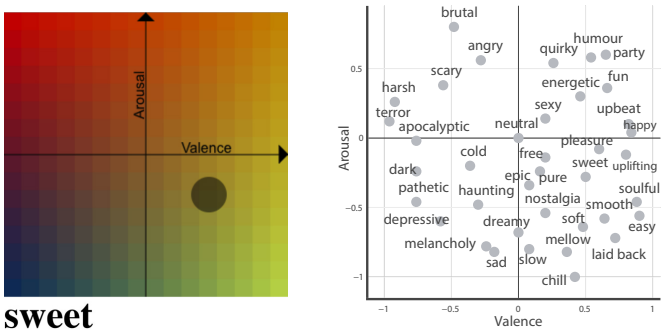
Participants annotated their perceived emotions using Mood Rater, a smartphone-friendly web application, whilst listening to the concert. Mood Rater was originally developed for the Mood Conductor framework [39] for participatory music performance and was adapted for this study. The interface (**Figure 1a**) is based on the valence-arousal (VA) space derived from the Circumplex Model of Affect [40]. The model proposes that most affective states can be associated with this two-dimensional space. The valence dimension describes how positive or negative an emotion is, while the arousal dimension characterises the level of excitation. The space's quadrants (Q) refer to emotions sharing similar characteristics: Q1 describes energetic positive emotions like "happy" and "fun," while Q2 describes energetic yet negative emotions, such as "angry" or "scary." Q3 comprises low energy and negative feelings like "depressive" and "sad," and Q4 low energy yet positive emotions such as "mellow" and "laid back." The VA space is commonly used in cognition studies to provide quantitative measures of emotion by mapping responses to numerical coordinates in the space.

The Mood Rater interface displays emotion tags; when tapping a specific point, the emotion tag closest to the selected coordinate, such as "sweet," (**Figure 1b**) appears. These tags are curated editorial tags extracted from I Like Music's (ILM)[3] collection, mapped to locations in the VA

2. The performance can be found at: https://youtu.be/55JJLq3ewHss. The video's progress bar has been divided into the corresponding 45 sections for navigation to a specific performance segment.

3. https://web.ilikemusic.com

space [41]. Annotations on Mood Rater are time-stamped based on HTTP GET request times logged on the server side. Synchronisation of the annotations to the live performance was done with a reference signal (similar to a clapperboard) which can be identified in the audio-video recording, as well as the server log, through a synchronous HTTP request.



(a) (b)

Fig. 1: (a) Mood Rater's interface, as displayed on participants' mobile devices. (b) Associated guide tags in VA space.

2.3 Procedure

During the performance, audience members were invited to participate by reporting their perceived emotions using Mood Rater. The audience members were instructed on how to access Mood Rater from their personal devices. A brief overview of the VA space was also given, and participants were able to acquaint themselves with the interface and preview the tags by tapping around their screens in a test run before the performance. Participants were instructed to use the application when they perceived a change in emotion expression by tapping on the interface; audience participants were able to send a new rating at any time during the course of the performance. Participants were given freedom to annotate at their own discretion, with the hope that this would provide a view of the moments during the piece when participants perceived a change strong enough to warrant making a new emotion annotation without being prompted. Following the performance, participants provided their gender and age and reflected on the user experience.

2.4 Participants

Invitations to the performance were made through QMUL campus mailing lists. Audience members were then invited to participate in the study. 15 participated out of approximately 30 concert attendees in the chosen performance. Of these, 13 completed the post-task questionnaire; 6 male and 7 female, aged from 23 to 36 years (M = 26.8, SD = 3.8 years).

2.5 Results & Discussion

Over the course of the performance, 949 total emotion annotations were collected (**Figure 2**). The collected data points were nearly evenly spread over all VA quadrants; in Q1: 332 points (35% of all annotations made), Q2: 253 (27%), Q3: 158 (17%), Q4: 206 (23%). Although the concentration of points in Q1 suggests more energetic and positive emotions were perceived, this distribution supports the idea that the chosen musical work is shows as variety of emotions which contrast between movements. For example, the VA ratings in the softer and slower second movement largely occupy Q4, while those in the lively and rapid third movement are clustered in Q1. Compared to the mean rating of the full piece (Arousal: M = 0.55, SD = 0.24; Valence: M = 0.53, SD = 0.22), the mean varied between movements on both arousal and valence, as shown in **Table 2**. This suggests that perceived emotion varies at least across movements for this performance of the Babajanian trio, and indicates that a single emotion descriptor would not be sufficient to characterise the whole piece.

The mean number of ratings per participant was 66.4 for the whole piece (SD = 88.3). On average, participants made 2.76 ratings/minute; this ranged from 0.15 ratings/minute to 10.65 ratings/minute. This wide variance in annotations supports the idea that some listeners are more aware of fine emotion cues than others or may be more sensitive to particular musical features. Participants who did not rate as frequently may not have perceived sufficiently strong emotion changes to warrant making an annotation, or may have been more focused on the live performance.

Participant Agreement Over Time. Previous music emotion studies have adopted various measures of inter-rater reliability (IRR) for assessing the agreement of emotion ratings between different raters [42], [43], [44], [45], [46]. In this work, we used intra-class correlation (ICC) to assess the IRR among participants' emotion ratings; this was adapted to assess the consistency of rank variables (here, valence and arousal) across more than two raters [47]. Specifically, we used two-way mixed, consistency, average-measures of ICC, notated ICC(3, K), to measure the extent to which similar rank orders can be found across participant annotations. It is worth noting that ICC(3, K) is mathematically equivalent to Cronbach's α [48], which is commonly used in assessing internal consistency (reliability) of continuous emotion annotations [16], [45], [46]. ICCs at longer timescales (e.g. a movement or full piece) and with more items being tested can potentially be inflated [49], [50]. Therefore, the performance was broken down into 45 segments based on the rehearsal letters marked in the score to offset possible bias in the analysis. The segments last from 11 to 72 seconds (M = 31.7, SD = 15.8) in the recorded performance, with 16, 9, and 20 segments in the three movements, respectively.

The individual emotion ratings were re-sampled using a step function at 1 Hz (one rating/sec) for the ICC calculation, where a rating is assumed to be unchanged until a new rating is made. The sampling rate adequately captures the meaningful changes in participants' emotion annotations, as even the most actively rating participants made up to 10.65 ratings/minute in the Live study, which is well below one rating/sec. This assumption is in line with the instructions given to participants, to rate when a change is perceived[4]. **Figure 3** shows the ICC estimate with a 95% confident interval [51] in each of the 45 segments, as well as the number of ratings for each segment[5,6]. **Table 3** shows the number of

4. However, if the persistence of perceived emotions is assumed to be decreasing over time, other modelling could be applied, e.g. Gaussian process interpolation.

5. There were no ICC(3,k) estimates for Segment 32 because no ratings were made in this segment.

6. The y-axis range in the figure is set to [-1,1] for figure clarity.

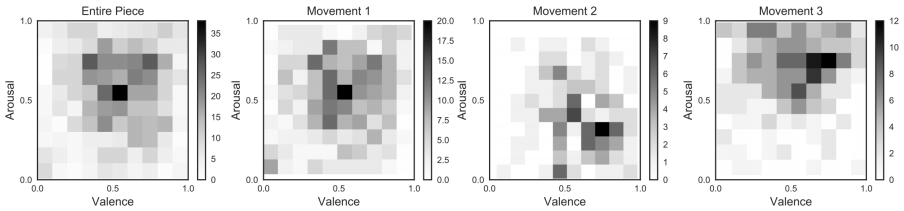


Fig. 2: VA rating distribution from the Live study through the piece (left subfigure) and in each movement (right subfigures).

M	Arousal Mean (SD)	Valence Mean (SD)
1	0.53 (0.22)	0.52 (0.23)
2	0.40 (0.22)	0.60 (0.19)
3	0.60 (0.20)	0.52 (0.23)

Tbl. 2: Live study mean and SD for all VA ratings on each movement (M).

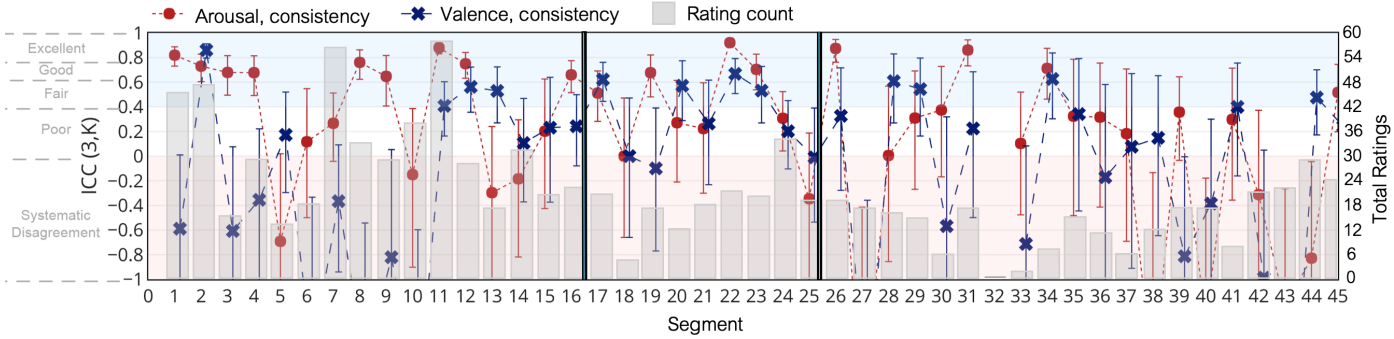


Fig. 3: ICC(3, K) estimates for each segment of the Live Study (at the 95% confidence interval) for arousal (red dots) and valence (blue crosses) ratings made in the Live study. Vertical grey bars indicate the total number of ratings made in each rehearsal segment; black vertical lines indicate the boundaries between movements.

segments associated with each level of agreement (excellent, good, fair, poor, systematic disagreement [52]) for VA.

Agreement Level	ICC(3, K)	Segment Count	
		Arousal	Valence
Excellent	[0.75, 1.00]	7	1
Good	[0.60, 0.74]	8	4
Fair	[0.40, 0.59]	2	8
Poor	[0.00, 0.39]	15	12
Systematic Disagreement	$[-\infty, 0.00]$	13	20

Tbl. 3: Cicchetti's Agreement Levels and respective ICC(3, K) ranges [52], with occurrences from Live study segments.

The agreement spanned the entire scale from systematic disagreement to excellent agreement for both the arousal (-2.09 < ICC(3, K) < 0.92, M = 0.15, SD = 0.71) and valence ratings (-2.53 < ICC(3, K) < 0.86, M = -0.14, SD = 0.8). **Figure 3** depicts the ICC(3,K) estimates for each segment of the Live Study, as well as the overall number of ratings made in each rehearsal segment. We observe that agreement changes quickly at the segment level, sometimes moving from near complete agreement to total systematic disagreement in consecutive sections. Several reasons may contribute to the low agreements; firstly, participants may perceive or rate the emotion at different timescales. Participants may also pay attention to non-performance factors that are less controlled in a live concert, such as audience noise or the actions of participants around them. Although participants were invited to explore the Mood Rater app after instructions were provided, there was no explicit trial of making ratings in context prior to the concert. Participants may not have understood the tags on the rating tool well; further, specific moments in a musical piece may have multiple contradictory or ambiguous emotion cues, making it difficult for listeners to perceive a singular emotion or select an appropriate rating to match this perception.

The reliability of time-continuous valence and arousal annotations collected in comparable studies also varies. For example, [46] reported very high internal consistency (Cronbach's α) of participants' ratings on arousal (0.998) and valence (0.998) on 794 clips, each annotated by at least 10 different listeners. Cronbach's α was also very high (>0.89) on both arousal and valence for all 8 pieces annotated by 52 listeners in [45]. In contrast, for the DEAM dataset [16], varied agreement levels were found for arousal ($0.28 < \alpha < 0.66$) and valence ($0.20 < \alpha < 0.51$); this aggregates emotion ratings on 1744 clips collected across three years, each annotated by 5 to 10 listeners. High agreement for annotation data may be partially explained by the choice of stimuli [45] and the selection of participants to ensure that consistent ratings are obtained (discarding disagreeing participants) [46]. Our results present varied agreement levels across segments within one piece among the same group of participants. In order to better understand such variability, a follow-up study was conducted with the aim of examining the rationale behind differing listener annotations, where listeners will be able to reflect on their time-based ratings retrospectively (see Section 3).

Emotion Rating Experience Feedback. The post-performance questionnaire collected participants' view of the overall ease of using the app, the difficulty levels of the rating task, and the impact of the guide tags. Each question was followed by an optional comment box for participants to leave further feedback. Lastly, an open-ended question "*Do you have any other suggestions on how we could improve our Mood Rater app?*" was presented. Age and gender were also collected from participants. The questions and corresponding responses to

the questionnaire are presented in Appendices A.2 and A.3.

Out of 13 participants, most participants (11) found the app 'easy' or 'very easy to use'. Over half of participants (7) evaluated the task of rating perceived emotions during the performance as 'easy' to 'very easy' while 2 participants found the task 'difficult;' 5 participants reported that the rating process distracted them from the performance while 3 reported no distraction; Over half the participants (7) evaluated the mood tags as 'useful' or 'very useful,' while 3 participants evaluated them as 'not useful.'

From the evaluation feedback, we can conclude that most participants considered Mood Rater overall to be successful in facilitating the self-reporting of real-time emotions conveyed by the music; however, the results highlight that such rating tasks tend to distract some of the participants from the actual performance. Open-ended feedback suggested improvements, especially in terms of emotion tags and interface design (see Appendix A.3). People with unfavourable opinions found the tags to be inaccurate and not adapted to the music, or felt they did not match their current emotion state. Mood Rater was consequently improved for further study, with revised mood tag choices and placements, and an updated interface to make it more engaging and understandable.

3 Lab Study: Reflective Rating Feedback

The Lab study further explored rehearsal segments found in the Live study to have varied agreement, with the aim to determine the reasons for the divergent ratings.

3.1 Music Performance Stimuli

The audio-video recordings of the Babajanian trio from the Live study performance were the stimuli for the Lab study. The first two movements (M1, M2) were chosen for perceived emotion annotation. These movements comprise the first 25 rehearsal segments (S1 - S25), which together last approximately 17 minutes[7]. In addition to annotating the first two movements, for seven segments (S5, S7, S12-14, S17) participants additionally reviewed and provided reasons for their emotion judgements. These excerpts were chosen based on the diversity of musical features, including varying instrumentation, dynamics, and tempo; in addition, these segments were determined to span a variety of agreement levels and VA emotion rating trends in the Live study. **Table 4** presents the ICCs for these seven segments, as calculated in the Live study. These ICC values range from -0.69 to 0.75 (M = 0.16, SD = 0.56) for arousal, and from -0.37 to 0.86 (M = 0.35, SD = 0.41) for valence.

3.2 Annotation Setup

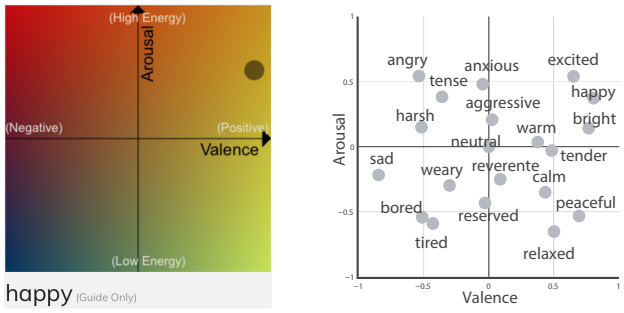
Participants made annotations via Mood Annotator, a web-based software adapted from Mood Rater for this study.

Emotion Rating Function. Mood Annotator enables time-varying emotion rating collection. The VA interface (**Figure 4a**) is positioned next to a window which displays the audio-video recording from the original Live study performance (**Figure 5**). Corresponding emotion tags included in the VA space were added to give participants a frame of reference as in the previous study, since a majority of participants found the tags useful (see Section 2.5). However, as some participants reported that the tags used in the Mood Rater app seemed inaccurate or confusing, we updated the tag choices and placements and explicitly indicated in the UI that these served only as guides in Mood Annotator. We improved the selection of tags based on previous work [53] which identified widely used music tags both from the music service AllMusic (AMG) and entries in the Affective Norms for English Words (ANEW) dataset [54]. Tags were selected based on the consistency of associated valence and arousal measures across raters in the ANEW dataset; tags with SD < 2.5 in either arousal or valence were considered to keep a balance between consistency and VA space coverage[8]. Following this process, 14 tags were selected from the AMG dataset which were relevant to the selected classical piece and avoided redundant meanings in the set. In addition, we selected another six tags from ANEW that were not included in the aforementioned AMG tags, but which we deemed important in the VA space interpretation ("calm", "happy", "bored", "neutral", "excited", "tired"). Each tag's location in the UI is represented by a closed disk, with the centre positioned on the ANEW average values and a diameter equal to the smallest Euclidean distance between any two out of the 20 tags (**Figure 4b**) on the VA space. For areas on the VA space not covered by emotion tags, no guide tag was presented.

Segment	ICC(3,K)	
	Arousal	Valence
S2 (M1)	0.73***	0.86***
S5 (M1)	-0.69	0.17
S7 (M1)	0.27*	-0.37
S12 (M1)	0.75***	0.56***
S13 (M1)	-0.3	0.53***
S14 (M1)	-0.18	0.11
S17 (M2)	0.51***	0.62***

Agreement
excellent
good
fair
poor
disagreement

Tbl. 4: ICC(3, K) for the 7 pre-selected segments from Live study annotations. Significance for the null hypothesis (ICC = 0): $p < .05$ (∗), $p < .01$ (∗∗), $p < .001$ (∗∗∗).



(a) (b)

Fig. 4: (a) Mood Annotator's interface with guide-only tags for the Emotion Rating Task. (b) Associated guide tags.

Emotion Reflection Function. Mood Annotator allowed participants to re-watch the recording and reflect on their

7. The cropped recording presented to participants in the Lab study can be found at: https://youtu.be/MHBfGm0SsYo. Timestamps included through the remainder of this paper reference this recording.

8. The ANEW ratings' SDs range from 0.31 to 3.38.

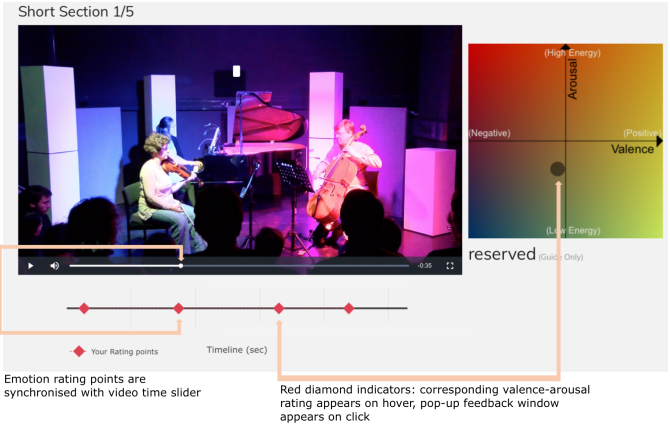


Fig. 5: Emotion Reflection Task; rating points (red diamonds) are not visible during the Emotion Rating Task.

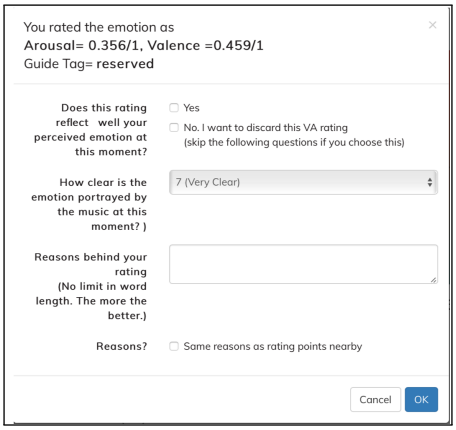


Fig. 6: Pop-up window displayed for reviewing and providing feedback for an example annotation point.

VA emotion ratings after an initial rating of the piece had been made. Listeners were presented with several short video recordings of the segments pre-selected from the whole music piece. For each segment, a timeline is included under its video to show where emotion rating points have been made, represented by red diamonds in **Figure 5**. When hovering on a timeline point, the participant's original emotion rating is simultaneously displayed on the VA space for reference. When clicking on a timeline point, a pop-up window (see **Figure 6**) is displayed for providing explanation feedback. Within this window, a participant can confirm or discard their previous rating. If the rating is confirmed, the participant is asked to select the clarity level of the emotion on a Likert scale from 1 (very unclear) to 7 (very clear). A comment box is further provided to allow participants to provide open-ended "Reasons behind your rating".

3.3 Procedure

Participants annotated on a 13" MacBook Air in a sound proof listening studio at QMUL. Audio stimuli were presented through headphones (Beyerdynamic DT 770 Pro) and video on the laptop display. Participants were able to adjust the audio level to their comfort before the initial task. Participants were given a brief overview of the VA space and the annotation software and explored the tag placement mapping on the VA space, as was done in the Live study. Participants were given time to acquaint themselves with the software during a trial. Once confident with the annotation procedure, they completed the **Emotion Rating Task** by annotating their perceived emotion in the VA space through the first two movements of the Babajanian Trio, presented as audio-video recorded from the Live study.

After rating the full movements, the rating timeline became visible (**Figure 5**) and participants embarked on the **Reflection Task**. Participants provided reflective feedback for each of the seven pre-selected musical segments discussed in Section 3.1 sequentially, for musical continuity. Participants were asked to review their previous ratings to provide open-ended explanations for their annotations. Participants were informed that there were no right or wrong answers and were encouraged to provide as much information as possible.

After finishing the Reflection task, participants completed the Goldsmiths Music Sophistication Index (Gold-MSI) [55] to determine their relative level of music experience and basic demographics. This background information was collected, in comparison to the limited information collected in the Live study, in order to examine whether musical experience could explain different emotion perceptions in listeners. The study duration ranged from 1.5 to 2.5 hours.

3.4 Participants

A new group of 21 participants (11 male, 10 female), distinct from that in the Live study, was recruited through an open call on the QMUL campus mailing list. All but one participant completed the full study.[9] Group scores for four sub-factors of the Gold-MSI are reported in **Table 5**. A majority of participants had at least 10 years of musical experience and was engaged in regular, daily practice of a music instrument (11), while the others had either novice to intermediate experience (5) or no musical experience (5). The ages ranged from 23 to 46 (M = 28.8, SD = 5.5). All participants were fluent English speakers and resident in the UK at the time of the study, and represented a variety of national backgrounds: 10 of the participants were Chinese, while the remaining 11 had Western backgrounds covering England (2), Greece (2) Spain (2), France (1), Italy (1), Germany (1), USA (1), and Costa Rica (1).

3.5 Results & Discussion

3176 VA emotion ratings were collected in the **Emotion Rating Task**. Similarly to the data collected in the Live study (see Section 2.5), data collected in the Lab study also spanned all four quadrants of the VA space: Q1: 1263 annotations (39%), Q2: 834 (26%), Q3: 460 (16%), Q4: 614 (19%). Compared to the annotations made on the first two movements in the Live study (Q1: 30%, Q2: 22%, Q3: 20%, Q4: 27%), high arousal quadrants (Q1, Q2) received proportionally more annotations than low arousal quadrants (Q3, Q4) in the Lab study. In the **Emotion Reflection Task**, 21 participants re-evaluated the 1098 VA ratings they gave for the seven pre-selected segments. 8 participants discarded 23 previous ratings, and 7 participants provided 12 new ratings. A total of

9. One participant completed the initial emotion rating task but left due to personal reasons before completing the Reflection Task. The participant later completed and returned the Gold-MSI by email.

	Range	Mean	SD
Engagement	9 – 63	44.57	10.61
Perception	9 – 63	50.81	7.51
Training	7 – 49	31.14	12.33
Emotion	6 – 42	32.67	3.93

Tbl. 5: Lab study group scores for Gold-MSI sub-factors.

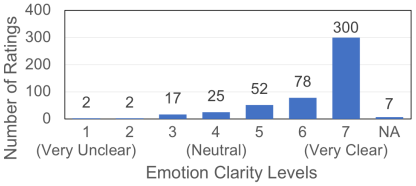


Fig. 7: Emotion clarity level bar graph.

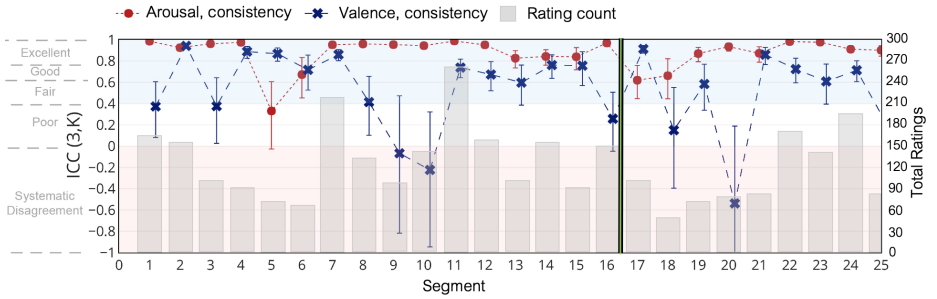


Fig. 8: ICC(3, K) estimates for each segment (95% confidence interval) for arousal (red dots) and valence (blue crosses) ratings made in the Lab study. Vertical grey bars indicate the total number of ratings made through each segment; a black vertical line indicates the boundary between movements.

483 VA ratings were accompanied by explanatory feedback, ranging from 2 to 24 comments per participant (M = 23, SD = 9), including 16 amended ratings with VA value changes, and 40 amended ratings with a time-stamp change. The rest of the re-evaluated ratings reported the same reasons as those for ratings close to them in time. **Figure 7** presents the emotion clarity levels reported by participants for their ratings (1 corresponds to *very unclear* and 7 to *very clear*, *NA* was used when participants chose not to rate the emotion clarity). The emotion clarity levels are not normally distributed, W(483) = 0.68, $p < .001$. Rather, the variable $X = 7-c$, where c denotes the emotion clarity level, seems to follow an exponential distribution ($1 \leq c \leq 7$). The results indicate that participants have reported most of the VA ratings with confidence, with the average clarity being 6.27 out of 7 (SD = 1.16) and the median being 7. The open-ended reasons behind the ratings contained over 7000 words, on which we undertook thematic analysis, as described in Section 4.

Rating Frequency. The number of annotations per participant ranged from 60 to 396 for the piece (M = 151, SD = 96) in the Lab study. Like in the Live study, the rating frequency in the Lab study varied across participants, ranging from 3.67 ratings/minute to 22.26 ratings/minute, with an average of 8.76 ratings/minute. Notably, the average participant in the Lab study rated nearly two times more frequently than in the Live study for the first two movements (4.21 ratings/minute on average in the Live study, t(34) = 3.72, $p < .001$).

There are some possible reasons for this difference: this participant group agreed and registered in advance to take part in an organised lab study, so may have been more focused and prepared, compared to the audience members who volunteered at a live concert; during the Live study, social factors could have limited ratings: participants may have been more hesitant to annotate, not wanting to distract the other audience members by using their phones. The setting and the socio-cultural norms dictated in a performance venue are also likely to impact the emotion perception. In the Lab study, more relevant guide tags were used in the rating tool, and the capability to pause and rewind the video recordings to re-visit and reflect on their annotations likely helped participants to make more confident ratings. Future reproductions of this study in different live venues would be beneficial to understanding the full impact of the performance setting on emotion perception.

Participant Agreement Over Time. As done in the Live study (Section 2.5), ICC values were computed at the segment level for participants' VA ratings. The ICC estimate with a 95% confident interval in each of the 25 segments and the number of ratings for each segment in the lab study are presented in **Figure 8**. The resulting ICCs for arousal ratings indicate good-to-excellent agreement for a majority of segments (0.33 $<$ ICC(3,K) $<$ 0.98, M = 0.87, SD = 0.15), with the exception of fair agreement for S5. For valence ratings, the agreement level varies from systematic disagreement to good agreement (-0.53 $<$ ICC(3,K) $<$ 0.94, M = 0.52, SD = 0.15).

There was comparatively higher agreement in the Lab study than in the Live study for the 25 tested segments. The ICCs of arousal ratings in the Lab study (M = 0.87, SD = 0.15) compared to those in the Live study (M = 0.37, SD = 0.44) demonstrated significantly stronger agreement, t(24) = 6.91, $p < .001$. The ICCs of valence ratings in the Lab study (M = 0.52, SD = 0.38) are also significantly higher than those in the Live study (M = -0.08, SD = 0.76), t(24) = 4.18, $p < .001$).

The greater agreement levels may be attributed to differences in the listening conditions. Participants might have had a better understanding of the emotion rating task and greater focus and concentration in the lab setting. Each made and reflected on their annotations alone, with few distractions compared to a real-world live performance setting involving audience etiquette and social interactions, as previously mentioned when discussing rating frequency. Furthermore, the ability to replay the Lab study recording and reevaluate annotation points may have allowed participants to refine and clarify their judgements, leading to more agreement than that of judgements based on listening to only one play-through. Although there are clear differences between the two studies, the ICC values of the Lab study reinforce the observations from the Live study, indicating that there are widely varying levels of agreement through the piece. Some sections, such as S9 and S10, display consistent systematic disagreement of valence, regardless of setting. Despite listening to the exact same performance, there are many differences in emotion perception between the two studies, further emphasising the need to address the underlying cognitive interactions and reasoning behind the annotations.

4 Thematic Analysis of Reflections

Further analysis aims to address these rating inconsistencies by focusing on the underlying causes for the disparate emotion judgements.

4.1 Methodology

Two of the authors (SY, CNR) jointly conducted an inductive "bottom-up" thematic analysis [32], [56] of the participants' feedback (Section 3.5). The comments made by participants were annotated by each author independently in NVIVO12[10]. Each author first generated a series of "codes" that identified notable semantic or latent content in the feedback. Independently identified codes and themes were then refined through a joint discussion of overlaps and divergences. A final set of themes emerged, which both researchers concur were notable and reliable for further joint iterative coding. Both researchers performed the final round of thematic coding by categorising each comment to fit within one or more codes of the final set of emerging themes.

4.2 Emergent Themes of Emotion Perception

Seven key themes motivating perceived emotion annotations appear in participants' open-ended reflections: *(1)* Perceptual Acoustic Features, *(2)* Instrumentation & Arrangement, *(3)* Personification of Instruments, *(4)* Expectation & Violation, *(5)* Musical Structures, *(6)* Performer Expression, and *(7)* Stage & Visuals.

4.2.1 Perceptual Acoustic Features

This theme includes material about music characteristics, the most commonly referenced codes. As participants made emotion judgements in a time-based manner, this theme can include both elements of *feature quality* and *feature variation*.

Feature quality involves music features arising at the time of annotation. The codes found include musical features such as **melody, timbre, timing, harmony**, and **dynamics/loudness**. The importance of these features to emotion perception in Western tonal music are described in [2], [11], [57], and noted by participants:

> *"Violin only, timbre bright; high pitch leads to a high valence feeling. slow tempo and relatively low loudness lead to low valence..."* - P7, S5 (11:25)

Feature variation refers to comments on changes in these musical features, the evolution of which tells us about the changing emotion of a performance. Codes include **dynamic change, harmonic progression, melodic progression, timing variation**, and **timbre variation**, each of which are common foci in the modelling of music emotion [5], [16].

> *"...a transition point in the music where we are moving to a more positive and hopeful place. This is evident in the increased brightness of the sound and change of tempo to that of [sic] a little faster."* - P14, S2 (01:40)

Certain features, such as timbre [58], are described as sound qualities at given time points while other features such as dynamics and harmony, are more commonly discussed in the context of their changes. This theme affirms the results of prior studies [12] and highlights the need for MER models to examine features not only at defined points in time but also as trajectories shaped through time.

10. https://www.qsrinternational.com/nvivo/nvivo-products

4.2.2 Instrumentation & Arrangement

Performers can control and shape timbral features specific to their instrument, an important factor for expressiveness [58]. This theme involves attending to the sounds of an instrument or the particular instrumental arrangement and is related to *Perceptual Acoustic Features*. Participants frequently referred to the three instruments of the trio – **violin, piano** and **cello** – and their interactions.

In the Western canon, different instruments assume distinct roles in an ensemble. Participants remarked on this, yielding the **lead instrument** and **instrument interaction** codes. When instruments are playing solo or carrying the melody, participants' comments suggest that these lead instruments are responsible for conveying the emotion while the supporting instruments provide context. The violin in this piece often acts as a lead instrument, while the piano and cello generally provide the accompaniment; however, each of these instruments occasionally take the leading role. At moments when the hierarchy of the instruments change, the annotations frequently describe which of the instruments moves into the lead position and how the others respond:

> *"...starts with a solo piano that slowly picks up in tempo and volume. The cello and violin respond to the theme presented by the piano which leads to a conversation between the instruments that joins together in the end with higher energy."* - P19, S3 (05:11)

Participants describe the interaction between instruments as they come in and out of active playing. For example, participants note when multiple instruments play the melody in unison and describe different instruments as working together to convey an emotion:

> *"when the violin is alone it sound [sic] sad, like alone. when the cello start the [sic] go together and sounds more positive."* - P21, S3 (07:50)

Instruments' sound character and variations will therefore influence the perception of emotion, highlighting the need for further exploration of these aspects in MER research.

4.2.3 Personification of Instruments

This theme presents a novel insight, covering comments that describe instruments communicating emotion like in human-to-human interaction and the use of abstract metaphorical language in music emotion perception. Participants' comments suggest they perceive emotion communicated by the instrument in the same way they would perceive emotion communicated by a person, suggesting a personification of the instrument itself. Participants may associate the sound quality and emotion of an instrument to human vocalisation; this capability to mimic the voice and express emotion through an instrument is well noted in the design and aesthetics of particular instruments [59], [60]. The sound can thus be described through evocations of images by the listener:

> *"The violin plays really long notes which resembles [sic] a wailing voice."* - P19, S17 (11:44)

> *"Dark timbre, sad melody, sounds like somebody is crying."* - P3, S2 (01:77)

Moreover, instead of making self-reflections such as "I perceive sadness," or "I sense agitation," listeners tend to attribute sadness or agitation to the instruments themselves:

"Lonely piano (playing by its own) playing a sad tone." - P8, S2 (01:20)

Instruments are described as "blue," "playful," or "romantic." This mirrors metaphorical language found in human communication, where complex concepts such as emotion are found in language via the use of metaphor; in this case, via embodiment and personification of inanimate objects [61]. In the perception and understanding of musical emotion, participants may view instruments as "living" beings which want to communicate and share their emotions.

4.2.4 Expectation & Violation

Musical play with expectations may lead to the induction of specific emotions [62]. The listener may experience pleasure when expectations are satisfied, or surprise when they are violated. When the listener does not hear what is expected, they may sense uncertainty and insecurity, as has been observed with temporal violations in musical performance [10].

"...there is a real tension which causes an anxious emotion to be perceived. There is a slight hesitancy to the piano part with a small delay in the playing of some of the notes which clashes with the fluidity of the strings. It isn't discordant, but causes a sense of anxiety about what is next in the music." - P14, S2 (03:09)

Participants are especially sensitive to portions of music that defy harmonic expectations, and the inherent tension that comes with note clashes. They feel unsure of what is to come and react to the instability in these changes:

"Chromatic movement makes me feel like something is about to change, although it is not yet very negative feeling... cello increases the loudness and the progression is very unexpected; it is hard to tell where the piece will go next." - P10, S4 (09:49)

Theories on expectation may be difficult to confirm, as listeners differ in their perceived expectations [63], but we see in this theme an immediate relevance to emotion perception.

4.2.5 Musical Structures

This theme includes comments surrounding the compositional structure or **musical form**. Given a piece of music, listeners may divide a sequence of sounds into segments and group these segments into categories [64]. Emotion changes are sometimes perceived at **boundaries and transitions** when sound states change or new material is introduced:

"A new passage starts here. However, it [sic] my ratings are not that clear, there is an increase in intensity from the trio, but the piano starts and [sic] alternating pattern in the bass that increases anxiety, maybe arousal should be increasing but valence should be lower (0.4)? " - P11, S14 (09:52)[11]

"Transition to the next state. From slow/sad music to more high energy still kind of sad music." - P8, S7 (05:02)

Repetitions of thematic material and motifs are fundamental to musical perception and provide listeners with a pattern of expectation, thus influencing perceived emotion intensity [65], [66]. Repetitions can also lead to the association and recall of imagery, as noted in *Personification of Instruments*:

"In this section the same theme is repeated with rising volume and confidence. I associate this pattern with images like: sunrise, rebirth or a new dawn which all have a positive, energetic connotation." - P19, S7 (04:16)

"... there is an ascendent [sic] repeating intervalic pattern that moves in arousal but keeps the joyful character." - P11, S14 (09:13)

4.2.6 Performer Expression

This theme refers to comments regarding how the performers impact the music they create and its emotion content. Musicians may alter the quality of a music note through its timbre and colour, articulation, and movement. Some participants associated the acoustic variations with **performance technique**, such as vibrato in the violin part or grace notes and arpeggios in the piano. Instrument-specific comments are mentioned by people with over ten years musical training in violin and piano respectively, and it indicates that people might pay more attention to the instrument they have expertise in playing for emotion perception. Different forms of **articulation**, for instance legato and staccato styles, are also reported. Examples include:

"... more violent rhythmic passage. There is a marked staccato." - P11, S12 (08:35)

"Vibrato, high arousal" - P7, S14 (09:40)

In addition, comments about **embodied expression**, refer to use of performer gesture and facial expression to convey feeling. Facial expression and body movements are known to convey emotion in a performance [67], [68], as well as information about the musical structure of a piece [69]. Participants referred to gestures such as bow movement and performers' facial expressions:

"...getting louder and more dissonant, cellist face looks very expressive, face screws up." - P13, S7 (05:11)

Performers ultimately provide the direct line of communication from the musical score to the audience. Individual interpretations can thus change the emotion quality of a piece, making gesture, and stage expression in the context of live music important aspects for MER. A piece performed by two different soloists will not sound exactly the same, nor will it have the same emotion nuances across performers or even individual performances.

4.2.7 Stage and Visuals

Two reactions were derived from visual reference, although there was no instruction in the task to examine beyond the music in the audio-video recording. Both refer to **stage lighting**; in particular, participants associated the decrease of arousal with the lights darkened in the final segment:

"The violin changes the length of the notes and with that the energy of the music. Also as it does not have light it is less energy as in the other parts." - P21, S17 (11:47)

Although this is not a frequently occurring theme, it does highlight the impact of context and setting. Even in a lab viewing of performance video, participants were able to associate emotion with changes in stage lighting, underscoring the relevance of environment on perception. As emotion contagion means members of an audience are likely to react in similar ways to their fellow concert-goers [70], [71], it is worth exploring the impact of environment, staging, and venue on the perceived emotion content for a performance.

11. Note that this comment is associated with a fairly low emotion clarity level of 4; low emotion clarity levels were rare compared to the generally very high clarity levels (see **Figure 7**).

Theme	Relevant Codes	CO	CO Per Capita (Mean (SD))		Sig. (p)	Cohen's d
			Musicians	Non-musician		
Perceptual Acoustic Features	**Feature Variation**	**367**	**22.17(10.68)**	**11.22(5.85)**	**.01****	**1.22**
	Dynamic Change	***130***	**8.00(5.51)**	**3.78(2.86)**	**.04***	**0.92**
	Harmonic Progression	***94***	**7.42(5.02)**	**0.56(0.68)**	**.00*****	**1.78**
	Melodic Progression	*69*	3.33(2.66)	3.22(2.44)	.93	0.04
	Timing Variation	*60*	2.58(2.10)	3.22(3.08)	.62	-0.25
	Timbre Variation	*14*	0.83(1.14)	0.44(0.68)	.37	0.40
	Feature Quality	315	17.33(5.37)	11.89(6.76)	.08	0.91
	Dynamics & Loudness	*40*	2.33(1.65)	1.33(1.25)	.15	0.67
	Harmony	***56***	**4.42(4.68)**	**0.33(0.67)**	**.01****	**1.14**
	Melody	*98*	4.58(2.22)	4.78(5.09)	.92	-0.05
	Timing	*60*	2.83(2.07)	2.89(2.08)	.95	-0.03
	Timbre & Roughness	*61*	3.17(1.82)	2.56(2.63)	.58	0.28
	Total	682	39.50(15.01)	23.11(11.16)	.01	1.21
Instrumentation & Arrangement	Instrumentation	235	11.92(6.58)	10.22(9.00)	.66	0.22
	Cello	48	2.17(2.03)	2.44(2.45)	.80	-0.13
	Piano	91	5.08(3.07)	3.33(2.98)	.23	0.58
	Violin	96	4.67(2.81)	4.44(4.14)	.90	0.06
	Instrument Interaction	63	3.08(1.38)	2.89(2.33)	.84	0.11
	Lead Instrument	39	2.33(1.43)	1.22(1.69)	.15	0.72
	Total	337	17.33(8.09)	14.33(12.54)	.56	0.29
Personification of Instruments		72	3.92(3.09)	2.78(3.08)	.44	0.37
Musical Structures	Boundaries of Sections	28	1.58(1.66)	1.00(1.15)	.38	0.40
	Repetition	**26**	**2.00(1.63)**	**0.22(0.42)**	**.00*****	**1.40**
	Transition	12	0.50(1.19)	0.67(1.05)	.75	-0.15
	Total	66	4.08(3.12)	1.89(1.97)	.08	0.81
Expectation & Violation		**64**	**4.25(3.39)**	**1.44(1.07)**	**.02***	**1.05**
Performer Expression	Embodied Expression	5	0.42(1.38)	0.00(0.00)	.34	0.40
	Articulations	**16**	**1.33(1.11)**	**0.00(0.00)**	**.00****	**1.59**
	Music Playing Techniques	11	0.83(1.14)	0.11(0.31)	.07	0.81
	Total	**32**	**2.58(1.93)**	**0.11(0.31)**	**.00****	**1.66**
Stage & Visuals		2	0.08(0.28)	0.11(0.31)	.84	-0.09
All		**1255**	**70.01(19.99)**	**46.00(23.43)**	**.02***	**1.24**

Tbl. 6: Resultant themes from analysis of participant annotation explanations. Group mean and SD of code occurrence (CO) by capita and significant differences are indicated, $p < .05$ ($*$), $p < .01$ ($**$), $p < .001$ ($***$).

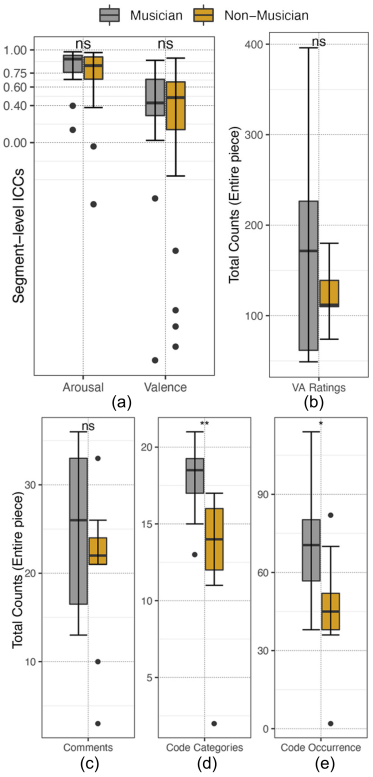


Fig. 9: Musicians' and non-musicians' (a) VA ICC agreement in 25 segments, (b) VA annotation counts, (c) explanation comments, (d) code categories, and (d) CO per participant in each group.

4.3 Code Occurrence (CO)

The number of code occurrences (CO) for any individual code has a potential maximum value of 483, the total number of comments collected in the study. **Table 6** presents the main themes along with their relevant codes and CO; as each overarching theme can include several codes, the CO number for a theme can exceed 483, as in the case of *Perceptual Acoustics Features*. It should be noted that some pieces of coded material are shared between multiple themes, as the underlying ideas in a comment can be related to multiple themes; for instance, *Personification of Instruments* and *Instrumentation & Arrangement* both include comments made about specific instrument parts and their levels of activity in the music. Although a CO is presented alongside each theme, it is important to bear in mind that themes with lower CO are not inherently less relevant [32]. The goal of thematic analysis is to gather together common threads within the participant feedback, and understand which points were sufficiently noteworthy as to be cited as emotion cues. In order to account for difference in group sizes (musician vs non-musician), a "code occurrence per capita" was determined by averaging the CO for each group. This CO/capita appears in **Table 6** next to each relevant code within the emergent theme.

4.4 Comparing Musicians' and Non-Musicians' Emotion Perceptions

Comparisons were conducted between music expertise groups for different aspects of emotion perception. Participants having 10+ years of musical training and scoring above average on the Gold-MSI were classed as "musicians" (N = 12); others are labelled "non-musicians" (N = 9).

Agreement Levels, and Rating and Comment Frequency. We first compared the agreement levels amongst the members of each group with ICCs of VA ratings in the 25 rehearsal segments. No significant differences were found in either arousal or valence rating agreement, as indicated by **Figure 9a**. The ICCs of arousal ratings from the Live study (M = 0.37, SD = 0.44, see Section 2.5) demonstrated significantly lower agreement than those from the Lab study, in both the musician (M = 0.82, SD = 0.19, t(33) = -4.74, $p < .0001$) and non-musician groups (M = 0.71, SD = 0.36, t(46) = -3.02, $p = .0041$). For valence ratings, the Lab study's ICCs of the musician group (M = 0.34, SD = 0.65) are significantly higher than those in the Live study (M = -0.08, SD = 0.76, t(46) = -2.11, $p < .04$). This was not the case for the valence ratings for the non-musician group (M = 0.12, SD = 0.92).

Rating and comment frequencies were compared using the number of VA ratings made per participant for the whole piece (**Figure 9b**), and the number of VA ratings accompanied by explanation feedback per participant in each group (**Figure 9c**). Although the group differences in rating and comment frequencies were not significant ($p = .16$, $p = .24$, respectively), musician participants, on average, made more VA ratings (M = 174.50, SD = 120.81) than non-musicians (M = 120.11, SD = 32.70), and provided more explanatory comments (M = 25.08, SD = 9.02) than non-musicians (M = 20.33, SD = 8.83). There is also greater

variance among the musicians, as shown by larger SD and inter-quartile ranges for these factors (**Figure 9b, Figure 9c**).

Code Categories and Code Occurrence. Additionally, we compared the number of code categories and the overall number of CO/capita for comments by both the musician and non-musician groups. This explored the difference in specific musical elements noted by members in each group and the range of reasons behind the annotations made. This was done to determine if the musical background of the participants may have contributed to the perception of different musical qualities.

Comments from the musician group yielded significantly more code categories (M = 17.92, SD = 2.27) than the non-musician group (M = 12.78, SD = 4.55), t(11) = 3.10, $p = .01$, as shown in **Figure 9d**. Also, the musician group's feedback mentioned significantly more codes (higher CO per capita) in general (M = 70.01, SD = 19.99) than that of the non-musician group (M = 46.00, SD = 22.43), t(16) = 2.55, $p = .02$, as shown in **Figure 9e**.

For individual codes, musicians' feedback on annotations included significantly greater CO in three themes: Perceptual Music Features, t(19) = 2.73, $p = .0099$, Expectation & Violation, t(14) = 2.57, $p = .02$, and Performer Expression, t(12) = 4.16, $p = .001$. Because of relevant musical experience, we would expect musicians' feedback to contribute the majority of the codes. Musicians are inherently be more atuned to the musical features; they also possess the training to identify and vocabulary to describe specific musical properties within a piece. In particular, musicians commented more often on harmony, t(12) = 2.85, $p = .015$, harmonic progression, t(12) = 4.47, $p = .0008$, and dynamic changes, t(17) = 2.17, $p = .04$. They report on *Feature Variations* more than non-musicians, t(18) = 2.86, $p = .01$. This supports existing findings that musicians focus more on harmonic movement than non-musicians [72], [73].

Musicians' feedback included significantly more references to *Repetition*, t(13) = 3.46, $p = .004$. This may be attributed to a number of factors, primarily that musicians are more aware of musical structure. Musicians may also be more likely to recognise repetitions or link certain phrases with an emotion. Within the theme of *Performer Expression*, musicians commented more often on *Articulation*, t(11) = 4.00, $p = .002$. They may have commented more on the performers' expressive actions because they themselves play an instrument and are more aware of the significance of the performer's actions. Awareness of different articulation styles and their connotations, particularly on different instruments, may again depend on musical knowledge and performance experience.

It has been previously established that musical training has significant effect on neural activations relating to emotion and reward while listening to music [10]. It is thus likely that musicians may listen with more empathy and focus compared to their non-musician counterparts.

The group differences of the other themes and codes were not significant ($p > .05$, see **Table 6**). As participants' agreement levels, rating frequency, and comment frequency do not differ between groups, non-musicians may be able to notice these particular features, but perhaps do not have the vocabulary or background knowledge to properly describe them.

5 Discussion

Here, we refer to three levels at which music can be observed: signal, perceptual, and semantic. Signal-based features refer to objective characteristics that can be directly computed or inferred from the music signal, whether in the audio or symbolic domain. Perceptual features concern musical properties that are subjectively evaluated by listeners and which depend on psychoacoustic factors. Semantic features are linked to the meaning that music conveys to listeners and can be characterised in linguistic terms [17], [74]. These different levels of features are interdependent and are found to be beneficial when combined in the modelling of emotion expression [17], [28].

Much progress has been made in MIR in recent years for automated signal-based feature extraction and feature learning using several inputs, including audio or symbolic data (such as MIDI, MusicXML, piano roll notation, and Music Encoding Initiative (MEI)). A range of feature extraction libraries and toolboxes have been proposed for both the audio [75], [76], [77], [78], [79], [80], [81], [82], [83] and symbolic [84], [85], [86], [87] domains. MER has been found to greatly benefit from these tools, and commonly relies on automated feature extraction [16], [88]. Still, investigations of links between perceptual and semantic features and these automated signal-based extractors are limited to a handful of studies [17], [28].

The creation of intuitive music emotion models requires a clearer understanding of how a listener perceives music emotion and how a machine can recreate this process. However, human emotion is rarely explicit in the way computers are; thus, an interdisciplinary exchange between MIR, music psychology, and musicology is necessary for a holistic view of music features and for more human-centred MER. Here, We present a general mapping between features at different levels and discuss gaps which can be addressed in future research, given the insights derived from our thematic analysis.

5.1 Connecting Emergent Perceptual Themes and MIR

We derived several perceptual and semantic features below according to each code from the thematic analysis (see Section 4.2), in reference to prior work [17], [28]. The semantic and perceptual features are linked to MIR features and automatic extraction tools as presented in **Table 7**. We focused on existing audio-based tools, namely MIRtoolbox [75], Vamp plugins[13], and the score-based tool jSymbolic [90]. Where appropriate, the need for additional theoretical or computational work is noted.

Existing signal-based methods cover many components of the *Perceptual Acoustic Features* theme. Features related to dynamics, timing, harmony, melody, can be extracted from both audio and symbolic data, while timbre correlates can be extracted from audio data. These features have been shown to be informative for music emotion modelling [109], [110], [111]. However, for several of the features, namely melodic movement, extraction approaches have yet to be established or remain to be incorporated effectively into MER systems.

12. https://github.com/bbcrd/bbc-vamp-plugins
13. https://www.vamp-plugins.org

Relevant Code	Perceptual and Semantic Feature	Representative MIR Feature	Feature Extraction by Existing Toolboxes		Related Computational Work Audio (A) or Symbolic (S)	Description
			Audio: MIRtoolbox [M] Vamp plugins [V-]	Symbolic: jSymbolic		
Perceptual Acoustic Features						
Dynamics	Loudness	RMS energy	mirrms [M] loudness [V-L]	–	Global sound energy [75], [89](A)	Volume/intensity, measured with global signal energy
	Dynamic change	Low energy rate	mirlowenergy [M] low energy ratio [V-B]	D-4 Average Note to Note Change in Dynamic	Low energy ratio [75](A), MIDI dynamic change [90](S)	Loudness contrasts, frames with less-than-average energy
Harmony	Mode	Mode	mirmode [M] key mode [V-Q]	P-33 Major/Minor	Modality estimation [75], [91](A), [90](S)	Overall mode: major, minor
	Chord	Chord type	Chordino [V-C]	C-3 Chord Type Histogram	Chord estimation [92](A), [90](S)	Type: major, minor, dominant, etc.
	Harmonic progression	Harmonic change detection	mirhcdf [M] Chordino [V-C]	—	Harmonic change [92](A)	Change in harmonic progression
	Key clarity	Key clarity	mirkeyclarity [M]	—	Key clarity [75](A)	Clarity of estimated tonal centre
Melody	Pitch	F0 estimate, MIDI pitch	mirpitch [M] fundamental freq. [V-L]	P-2 Pitch Class Histogram P-14 Mean Pitch	Pitch estimation [75], [89](A), [90](S)	Perceived pitch
	Melodic progresson	Pitch variability	—	P-24 Pitch Variability P-25 Pitch Class Variability	Pitch contour [93](A), Pitch variability [90](S)	Pitch increase/decrease
	Pitch range	Pitch value differences	—	P-8 Range	Pitch range [90](S)	Pitch range in semitones
	Inharmonicity	Inharmonicity	mirinharmonicity [M] inharmonicity [V-L]	—	Inharmonicity estimation [75], [89](A)	Degree of deviation of partials from harmonic series
Timing	Tempo	Tempo	mirtempo [M] tempo [V-Q]	RT-1 Initial Tempo RT-2 Mean Tempo	Tempo estimation [75], [91](A), [90](S)	Estimated tempo
	Tempo change	Tempo change	mirtempo [M]	RT-3 Tempo Variability	Tempo change [75](A), [90](S)	Tempo variation over time
	Note density	Note density or event density	mireventdensity [M]	RT-5 Note Density, R-10 Note Density per Quarter Note	Event density [75](A), [90](S)	Estimated note onset per second
Timbre & Roughness	Smoothness	Spectral flatness	mirflatness [M] spectral smoothness [V-L]	—	Flatness [75], [91](A)	Smoothness of the sound
	Dissonance	Roughness	mirroughness [M]	—	Roughness [75](A)	Dissonance of the sound
	Brightness	Spectral centroid/ rolloff	mirbrightness [M] spectral centroid [V-L]	—	Brightness [75], [91](A)	Brightness of the sound
Instrumentation & Arrangement						
Instrument	Instrument(s) present	Instrument recogniton	—	I-1 Pitched Instruments Present	Instrument recognition [94], [95], [96], [97](A), Instruments presented [90](S)	Which instruments are present
	Number of instruments	Number of instruments	—	I-8 Number of Pitched instruments	Musical layers distribution [98](A), Number of instrument presented [90](S)	Number of instruments present
Instrument interaction	Interaction	Layers/interaction	—	T-19 Parallel Motion T-21 Contrary Motion	Ratio of musical layers transition [98](A), Relations between independent voices [90]	Musical lines, interaction, entrances, active playing
Lead instrument	Lead/melody recognition	Prevalence/importance of single instrument	—	I-3 Note Prevalence of Pitched Instruments	Predominant instrument recognition [96](A), Instrument prevalence [90](S)	Instruments playing solo or having a lead melody
Personification of Instruments						
	Musical metaphor	Evocations/imagery	—	—	Mental image of sound [99](A)	Abstract metaphor/imagery used to relate sounds to emotions
Musical Structures						
Boundaries	Perceived boundaries	Segmentation/ grouping	mirsimatrix [M] Segmentino[V-S]	—	Music segmentation [75], [100], Melodic segmentation [101], [102](S)	Section definition (beginning/end)
Repetition	Repetition	Repeated motifs	—	—	Music loops extraction [103](A), Repeated theme and section [104](S)	Melodic patterns/repetitions and reproduced motifs
Transition	Section transition	Transition	mirnovelty [M]	—	Music transition [75](A)	Movement to new section/form
Expectation & Violation						
	Tension	Music tension	miremotion [M]	—	Music tension [105](A), [73], [106](S)	Rising intensity, impending climax
Performer Expression						
Articulations	Articulation	Envelope (ADSR)	mirattacktime [M]	RT-7 Average Time Between Attacks	Articulation [75](A), [90](S)	Flow of successive notes, eg. legato, staccato articulation
Techniques	Arpeggio	Arpeggio	—	M-8 Amount of Arpeggiation	Arpeggios pitch direction [107](A), Arpeggiation detection [90](S)	Chord is articulated through separate notes
	Grace note	Grace note	—	S-1 Number of Grace Notes	Grace note detection [90]	Stylistic embellishment through additional notes, eg. acciaccatura
	Vibrato	Vibrato	—	P-40 Vibrato Prevalence	Vibrato detection [108](A), [90](S)	Regular, pulsating change of pitch

Tbl. 7: Perceptual and semantic features identified and corresponding MIR features. Audio toolboxes include MIRtoolbox [M] [75] and Vamp plugins from QMUL [V-Q] [91], libxtract [V-L] [89], BBC plugins [V-B][12], Segmentino [V-S] [100], and Chordino [V-C] [92]; for symbolic, jSymbolic [90]. Related audio (A) or symbolic (S) computational works are also reported.

Previous work estimated possible melody contours based on 15 predefined patterns [93], and tested their use in emotion and genre classification [112], [113]. However, extraction of pitch progression and range remain a challenge: multipitch (multiple $f0$) estimation is considered to be one of the main challenges in current MIR research [114], [115], [116]. Compared to audio, data represented in symbolic format can provide an accurate estimation of features which rely on nominal information from the score, e.g. note density or average pitch. The main disadvantage of symbolic data is that the sonic properties of different instruments are lost, which is detrimental to recognition of expressive aspects such as timbre [58]. Considerably more digital music is available in audio than in symbolic representation. For example, improvisations recorded in audio which may be difficult or time-consuming to transcribe, as well as music which cannot be reflected accurately in Western notation, are also neglected when there is reliance on symbolic data alone. This limits the inclusion of certain musical styles in both MIR datasets and the MER systems derived from them.

Most of the codes found in the *Instrumentation & Arrangement* theme have associated features which can be extracted with jSymbolic. These features are mainly limited to symbolic data, which is currently a more reliable source for providing information at the instrument level, compared to mixed and

mastered audio data, from which it may be challenging to separate instrument stems [114]. Moreover, musical scores do not contain information on performers' interpretations. There is much computational work on instrument recognition [94], [95], [96], [97], but this has mainly been applied to solo instrument classification tasks rather than multi-label classification in polyphonic music. Novel audio-based music texture features proposed recently could help address this gap and have been applied to music emotion recognition [98].

The remaining results of this study outline new frontiers for MER; for instance, the *Arrangement & Instrumentation* theme has shown that listeners focus on instrument activity and interactions and view their roles as providing emotion cues. For real-time instrument recognition there are avenues to explore, specifically, better detection of the instruments playing at a given time, duration of instrument interactions, and instrument roles (lead/solo vs accompaniment), may benefit emotion prediction. This highlights the significance of work in audio source separation and supports previous findings that training on a multi-track dataset achieved better MER results [117][14]. With more multi-track datasets available for public use [118], [119], this link between arrangement, instrumentation, and emotion perception should be further explored.

Computational models of features relating to the other themes, *Personification of Instruments*, *Expectation & Violation*, *Musical Structures* and *Performer Expression* have been proposed, although audio or score-based computational extractors are not yet widely available for these features. The recognition of *Musical Structures* and their repetition through a performance would likely provide additional cues for MER systems [65], [66]. Related to *Personification of Instruments*, audio retrieval by sketching mental images of sound can be applied to the exploring of listeners' abstract emotion representations [99]. In *Expectation & Violation*, theoretical work on modelling tonal tension with symbolic data has been successful [73], [106], but only limited empirical tension retrieval work exists [120]. Because these features are perceptual, formalising, quantifing, and capturing their variations require further work that is still in its infancy. This leads to a lack of available datasets which are well-labelled and verified, thus resulting in less material for computational study. Also, the detection of particular higher-level features, such as embellishment and articulation, require multidimensional lower-level features, and thus their inclusion is also limited.

The focus on live performance brings to light the importance of relatively unexplored musical attributes, such as those in the *Performer Expression* theme. The relation between expressive features and emotions has been studied with regard to vibratos and articulations [121], [122]. Methods to characterise expressive techniques have recently been proposed such as detection of vibrato in violin and erhu [108], [123], arpeggios in multiple instruments [107], use of pedal in piano [124], and representative playing techniques in guitar [125], [126] and bamboo flute [127]. These embellishments are unique to individual performances and computational models would be useful for comparison of playing styles, performer identification, and individual instrument sounds [128]. Where performance or production videos are also available, features related to codes in *Stage and Visuals* involving embodied expression can contribute relevant cues. The present study demonstrates the potential importance of multimodal aspects from the performance space itself and proposes introducing other sensory material in MER systems besides auditory stimuli; namely visuals, given that many popular music streaming services include visual material. Multimodal emotion-sensing using computer vision [129] is therefore promising for the future design of music emotion studies, with more multimodal data exchange platforms and web applications merging to produce enriched music performance resources [130], [131].

14. We assume that some of the findings presented in this study, which focuses on contemporary classical music, would apply to other genres of music. However, it is likely that some of the listener-informed features would vary across genres–this requires further investigation.

5.2 Limitations

Several limitations should be considered regarding this research. First, the musical stimuli were limited to sections from one performance of a musical piece. Although the Babajanian Trio is well situated within the Western contemporary music canon and spans a wide range of characteristics, the findings drawn from this recording cannot be generalised to other music without further study. Second, although we deliberately chose this relatively unknown piece to avoid familiarity bias in emotion perception, it could be argued that a certain familiarity could be expected of a Western classical and contemporary music style, which has its own distinct set of expectations compared to other genres. Another limitation comes from the fact that our results are prone to low statistical power due to the small sample size; this makes the findings only suggestive. However, such a limitation is the cost of our attempt to extract more rich and accurate information from each individual's detailed annotations in the listening studies and provides the basis for examining notable emotion-engendering features in other music traditions. Future studies may compare participants' perceived emotions from different music traditions across a variety of musical selections with larger participant sample size.

Last, as multiple factors varied between the Live and Lab Study settings, the differences observed between the studies may be affected by factors other than the setting itself. However, as an exploratory study, the primary goal of the research was to better understand the connections between semantic features and the perceived emotion, rather than to provide a systematic comparison of live vs lab settings. The differences in agreement levels and rating frequency may be attributed to many differences in the listening conditions including but not limited to different participants' demographics, different stimulus length and presentation format, and different guide tags used in the rating tools (e.g. the improvements made in Mood Annotator based on participant feedback). Testing the effect of recorded lab vs live settings would require a controlled experimental protocol designed to that effect.

5.3 Conclusions & Future Work

Ultimately, because musical performance is a form of emotional communication, it will be perceived differently by individual listeners. We began by conducting a study

which collected time-based music emotion annotations from participants, then followed up with a study interrogating the reasons why participants consider certain musical points to be convey emotion. It is clear that emotion reactions to music are highly varied, as seen in both the VA agreement levels and also in the wide variance in the rate of annotations made by individual participants in any given section of the piece. We then examined listener music emotion perception based on high-level, salient emotion aspects identified in listener feedback. These features ranged from the qualities of the instruments themselves to the visual atmosphere and performer expression. Tasks in MER currently rely heavily on musical features related to tempo, dynamics, timbre, and melodic/harmonic progressions; although these musical elements are critical to emotion perception, they do not make up the full landscape of the emotion experience. Inclusion of more low-level acoustic features may result in more accurate models, but the ends do not always justify the means—these potentially confounding variables do not necessarily tell us about the underlying cognitive mechanisms at play and the subjectivity of data has been known to be problematic in large music emotion datasets. Our research suggests a way to accept such subjectivity as an inherent part of emotion, rather than as an "issue."

For future research on music and emotion, we advocate for providing more attention to high-level elements such as performer expression, instrumentation and interaction between instruments, musical structure, and visual elements of a performance environment. These elements matter to participants, and it is important that further research explores links between perception and cognition. In this sense, the idea of a performance being communication between composers, performers, and instruments and listeners is important; such processes involved in music emotion perception share similarities with that in interpersonal communication. This finding links MER with other types of emotion communication research and provides potential directions for further exploration between the communication sciences, computer science, and music cognition. In the development of MER tools, it is therefore important to consider the listener base. Those with relevant musical background may seek out different features to apprehend the emotion content than those without, and it may be beneficial to tailor systems individually to ensure consistency in emotion prediction. For instance, a music recommendation software targeted at classical music fans could ask the user some basic music experience questions during setup and incorporate this information in the analysis.

Additionally, as mentioned in Section 4.4, the role of language in perception of different musical elements is a prime area for future studies. It would benefit future work to acquire a better understanding of the way terminology is used to describe and relate to music, and the extent to which a knowledge of music vocabulary and concepts influences a participant's likelihood to attribute emotion to it. The Lab portion of this experiment could be run in two stages with non-musicians: following an initial rating as done here, the participants could be supplied with some basic music terminology and pedagogical rhetoric. Observations and discussions with participants on how their knowledge changes and how newfound language is applied in further reflections of the musical piece could help determine language's role in self-reports of music perception.

Finally, as mentioned in Section 5.1, visual factors from performers' body movements, stage, and lighting may also drive emotional perceptions. This indicates the audio-visual stimuli used in this study may express different emotions than stimuli limited to audio-only or visual-only information. Future research should also compare listeners' emotion perception with music stimuli involving different modalities. It would be worthwhile for emotion perception studies to explore the role of audiovisual stimuli and their interdependence.

Acknowledgments

SY is funded by a joint QMUL-CSC (China Scholarship Council) PhD scholarship. CNR is funded by an Electronic Engineering and Computer Science Principal Studentship from QMUL. EC is supported by an EU ERC H2020 Advanced Fellowship GA 788960 COSMOS. The authors would like to thank the reviewers and editors for their consideration and feedback, which have improved this manuscript.

References

[1] C. L. Krumhansl, "Music: A Link Between Cognition and Emotion," *Current Directions in Psychological Science*, vol. 11, no. 2, pp. 45–50, 2002.

[2] T. Eerola and J. 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.

[3] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1 suppl, pp. 123–147, 2002.

[4] R. Hiraga and N. Matsuda, "Graphical expression of the mood of music," in *Proc. ICMEW*. IEEE, 2004, pp. 2035–2038.

[5] E. Schubert, "Modeling Perceived Emotion with Continuous Musical Features," *Music Perception: An Interdisciplinary Journal*, vol. 21, no. 4, pp. 561–585, 2004.

[6] J. Akkermans, R. Schapiro, D. Müllensiefen, K. Jakubowski, D. Shanahan, D. Baker, V. Busch, K. Lothwesen, P. Elvers, T. Fischinger *et al.*, "Decoding emotions in expressive music performances: A multi-lab replication and extension study," *Cognition and Emotion*, vol. 33, no. 6, pp. 1099–1118, 2019.

[7] J. P. Bachorik, M. Bangert, P. Loui, K. Larke, J. Berger, R. Rowe, and G. Schlaug, "Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 26, no. 4, pp. 355–364, 2009.

[8] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "Emujoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[9] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller, "Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music." *Emotion*, vol. 7, no. 4, p. 774, 2007.

[10] H. Chapin, K. Jantzen, J. S. Kelso, F. Steinberg, and E. Large, "Dynamic Emotional and Neural Responses to Music Depend on Performance Expression and Listener Experience," *PloS ONE*, vol. 5, no. 12, p. e13812, 2020.

[11] A. Gabrielsson and E. Lindström, "The Role of Structure in the Musical Expression of Emotions," in *P. N. Juslin and J. A. Sloboda (Eds.), Series in Affective Science. Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2010, pp. 367–400.

[12] P. N. Juslin and E. Lindström, "Musical Expression of Emotions: Modelling Listeners' Judgements of Composed and Performed Features," *Music Analysis*, vol. 29, no. 1-3, pp. 334–364, 2010.

[13] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2005.

[14] E. M. Schmidt and Y. E. Kim, "Modeling Musical Emotion Dynamics with Conditional Random Fields," in *Proc. ISMIR*, 2011, pp. 777–782.

[15] V. Imbrasaitė, T. Baltrušaitis, and P. Robinson, "Emotion tracking in music using continuous conditional random fields and relative feature representation," in *Proc. ICMEW*. IEEE, 2013, pp. 1–6.

[16] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS ONE*, vol. 12, no. 3, p. e0173392, 2017.

[17] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1951–1963, 2014.

[18] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content-to context-based models," in *Proc. CMMR*. Springer, 2012, pp. 228–252.

[19] B. L. Sturm, "Evaluating music emotion recognition: Lessons from music genre recognition?" in *Proc. ICMEW*. IEEE, 2013, pp. 1–6.

[20] H. Terasawa, M. Slaney, and J. Berger, "The thirteen colors of timbre," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 323–326.

[21] K. Markov and T. Matsui, "Dynamic music emotion recognition using state-space models." in *MediaEval*. Citeseer, 2014.

[22] K. Cai, W. Yang, Y. Cheng, D. Yang, and X. Chen, "Pku-aipl'solution for mediaeval 2015 emotion in music task." in *MediaEval*. Citeseer, 2015.

[23] J.-J. Aucouturier and E. Bigand, "Mel Cepstrum & Ann Ova: The Difficult Dialog Between MIR and Music Cognition," in *Proc. ISMIR*, 2012, pp. 397–402.

[24] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception: An Interdisciplinary Journal*, vol. 27, no. 3, pp. 223–242, 2010.

[25] J. A. Sloboda, "Music Structure and Emotional Response: Some Empirical Findings," *Psychology of Music*, vol. 19, pp. 110–120, 1991.

[26] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. W. Schuller, R. Veltkamp, F. Weninger, and F. Wiering, "Emotional analysis of music: A comparison of methods," in *Proc. ACM Int. Conf. on Multimedia*, 2014, pp. 1161–1164.

[27] M. Schedl, E. Gómez, E. S. Trent, M. Tkalčič, H. Eghbal-Zadeh, and A. Martorell, "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.

[28] E. B. Lange and K. Frieler, "Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features," *Music Perception: An Interdisciplinary Journal*, vol. 36, no. 2, pp. 217–242, 2018.

[29] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing Emotion: An Overview," *Int. Journal of Synthetic Emotions (IJSE)*, vol. 3, no. 1, pp. 1–17, 2012.

[30] R. E. S. Panda, "Emotion-based analysis and classification of audio music," Ph.D. dissertation, 00500:: Universidade de Coimbra, 2019.

[31] A. Flexer and T. Grill, "The Problem of Limited Inter-rater Agreement in Modelling Music Similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[32] V. Braun and V. Clarke, "Thematic Analysis," in *PA Handbook of Research Methods in Psychology*, H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, Eds. Washington: American Psychological Association, 2012, vol. 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological.

[33] S. Yang, E. Chew, and M. Barthet, "Identifying listener-informed features for modeling time-varying emotion perception," in *Proc. CMMR*, 2019, pp. 438–449.

[34] H. Shoda and M. Adachi, "Why live recording sounds better: a case study of Schumann's *Träumerei*," *Frontiers in Psychology*, vol. 5, no. 1564, pp. 1–15, 2015.

[35] H. Shoda, M. Adachi, and T. Umeda, "How Live Performance Moves the Human Heart," *PloS ONE*, vol. 11, no. 4, p. e0154322, 2016.

[36] P. N. Juslin, S. Liljeström, D. Västfjäll, G. Barradas, and A. Silva, "An experience sampling study of emotional reactions to music: listener, music, and situation," *Emotion*, no. 8, p. 668, 2008.

[37] A. E. Greasley and A. Lamont, "Exploring engagement with music in everyday life using experience sampling methodology," *Musicae Scientiae*, vol. 15, no. 1, pp. 45–71, 2011.

[38] L. A. Warrenburg, "Choosing the right tune: A review of music stimuli used in emotion research," *Music Perception: An Interdisciplinary Journal*, vol. 37, no. 3, pp. 240–258, 2020.

[39] G. Fazekas, M. Barthet, and M. B. Sandler, "The Mood Conductor System: Audience and Performer Interaction Using Mobile Technology and Emotion Cues," in *Proc. CMMR*, 2013, pp. 15–18.

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

[41] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of musical mood: Comparison between crowd-sourced and curated editorial tags," in *Proc. ICMEW*. IEEE, 2013, pp. 1–6.

[42] X. Hu, S. J. Downie, C. Laurier, and M. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. ISMIR*, 2008, pp. 462–467.

[43] A. Aljanaki, F. Wiering, and R. Veltkamp, "Computational modeling of induced emotion using GEMS," in *Proc. ISMIR*, 2014, pp. 373–378.

[44] X. Hu and Y.-H. Yang, "A Study on Cross-cultural and Cross-dataset Generalizability of Music Mood Regression Models," in *Proc. ICM/SMC (Joint Conf.)*, 2014, pp. 1149–1155.

[45] N. Dibben, E. Coutinho, J. A. Vilar, and G. Estévez-Pérez, "Do Individual Differences Influence Moment-by-Moment Reports of Emotion Perceived in Music and Speech Prosody?" *Frontiers in Behavioral Neuroscience*, vol. 12, 2018.

[46] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The PMEmo Dataset for Music Emotion Recognition," in *Proc. 2018 ACM Int. Conf. on Multimedia Retrieval*, 2018, pp. 135–142.

[47] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological Bulletin*, vol. 86, no. 2, p. 420, 1979.

[48] L. J. Cronbach and R. J. Shavelson, "My current thoughts on coefficient alpha and successor procedures," *Educational and Psychological Measurement*, vol. 64, no. 3, pp. 391–418, 2004.

[49] J. M. Cortina, "What is coefficient alpha? an examination of theory and applications." *Journal of Applied Psychology*, vol. 78, no. 1, p. 98, 1993.

[50] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International Journal of Medical Education*, vol. 2, p. 53, 2011.

[51] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[52] D. V. Cicchetti, "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[53] Y.-H. Yang and J.-Y. Liu, "Quantitative Study of Music Listening Behavior in a Social and Affective Context," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1304–1315, 2013.

[54] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[55] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population," *PloS ONE*, vol. 9, no. 6, p. e101091, 2014.

[56] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[57] P. N. Juslin and J. Sloboda, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2011.

[58] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad, "Acoustical correlates of timbre and expressiveness in clarinet performance," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 2, pp. 135–154, 2010.

[59] P. N. Juslin, L. Harmat, and T. Eerola, "What makes music emotionally significant? Exploring the underlying mechanisms," *Psychology of Music*, vol. 42, no. 4, pp. 599–623, 2014.

[60] E. Schubert and J. Wolfe, "Voicelikeness of musical instruments: A literature review of acoustical, psychological and expressiveness perspectives," *Musicae Scientiae*, vol. 20, no. 2, pp. 248–262, 2016.

[61] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago: The University of Chicago Press, 1980.

[62] L. Meyer, *Emotion and meaning in music*. Chicago: University of Chicago Press, 1956.

[63] P. N. Juslin and D. Västfjäll, "Emotional Responses to Music: The Need to Consider Underlying Mechanisms," *Behavioral and Brain Sciences*, vol. 31, no. 5, pp. 559–575, 2008.

[64] J. Smith, "Explaining Listener Differences in the Perception of Musical Structure," Ph.D. dissertation, Queen Mary University of London, 2014.

[65] E. H. Margulis, "A Model of Melodic Expectation," *Music & Science*, vol. 22, no. 4, pp. 663–714, 2005.

[66] R. Simchy-Gross and E. H. Margulis, "The sound-to-music illusion: Repetition can musicalize nonspeech sounds," *Music & Science*, vol. 1, pp. 1–16, 2018.

[67] M. Leman and P. J. Maes, "The Role of Embodiment in the Perception of Music," *Empirical Musicology Review*, vol. 9, no. 3-4, pp. 236–246, 2014.

[68] J. W. Davidson, "Bodily movement and facial actions in expressive musical performance by solo and duo instrumentalists: Two distinctive case studies," *Psychology of Music*, vol. 40, no. 5, pp. 595–633, 2012.

[69] J. MacRitchie, S. Pullinger, N. Bailey, and G. Hair, "Communicating phrasing structure with multi- modal expressive techniques in piano performance," in *Proc. 2nd Int. Conf. on Music Communication Science*, 2009.

[70] E. Goffman, "The Neglected Situation," *American Anthropologist: Part 2: The Ethnography of Communications*, vol. 66, no. 6, pp. 133–136, 1964.

[71] A. J. Fridlund, "Sociality of Solitary Smiling: Potentiation by an Implicit Audience," *Journal of Personality and Social Psychology*, vol. 60, no. 2, pp. 229–240, 1991.

[72] E. Bigand, R. Parncutt, and F. Lerdahl, "Perception of Musical Tension in Short Chord Sequences: The Influence of Harmonic Function, Sensory Dissonance, Horizontal Motion, and Musical Training," *Perception & Psychophysics*, vol. 58, no. 1, pp. 125–141, 1996.

[73] M. M. Farbood, "A parametric, temporal model of musical tension," *Music Perception: An Interdisciplinary Journal*, vol. 29, no. 4, pp. 387–428, 2012.

[74] M. Buccoli, "Linking Signal and Semantic Representations of Musical Content for Music Information Retrieval," Ph.D. dissertation, Polytechnic University of Milan, Italy, 2017.

[75] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. DAFX*, 2007, pp. 237–244.

[76] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra *et al.*, "Essentia: An audio analysis library for music information retrieval," in *Proc. ISMIR*, 2013.

[77] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Science Conf.*, vol. 8, 2015.

[78] F. Eyben and B. Schuller, "opensmile:): the munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

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

[80] D. Cabrera, S. Ferguson, F. Rizwi, and E. Schubert, "Psysound3: a program for the analysis of sound recordings," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3247, 2008.

[81] C. McKay, I. Fujinaga, and P. Depalle, "jaudio: A feature extraction library," in *Proc. ISMIR*, 2005, pp. 600–3.

[82] C. Cannam, M. O. Jewell, C. Rhodes, M. Sandler, and M. d'Inverno, "Linked Data And You: Bringing music research software into the Semantic Web," *Journal of New Music Research*, vol. 39, no. 4, pp. 313–325, 2010.

[83] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, Firenze, Italy, October 2010, pp. 1467–1468.

[84] C. McKay, J. Cumming, and I. Fujinaga, "JSYMBOLIC 2.2: Extracting Features from Symbolic Music for use in Musicological and MIR Research," in *Proc. ISMIR*, 2018, pp. 348–354.

[85] M. S. Cuthbert and C. Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," 2010.

[86] T. Eerola and P. Toiviainen, "Midi toolbox: Matlab tools for music research," 2004.

[87] D. Müllensiefen, "Fantastic: Feature analysis technology accessing statistics (in a corpus): Technical report v1," *London: Goldsmiths, University of London*, pp. 140–144, 2009.

[88] Y.-H. Yang and H. H. Chen, "Machine Recognition of Music Emotion: A Review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 40, 2012.

[89] J. Bullock, "Libxtract: a lightweight library for audio feature extraction." in *Proc. ICMC*, 2007.

[90] C. McKay, J. Cumming, and I. Fujinaga, "JSYMBOLIC 2.2: Extracting Features from Symbolic Music for use in Musicological and MIR Research," in *Proc. ISMIR*, 2018, pp. 348–354.

[91] C. Cannam, M. Mauch, M. E. P. Davies, S. Dixon, C. Landone, K. C. Noland, M. Levy, M. Zanoni, D. Stowell, and L. A. Figueira, "Music information retrieval evaluation exchange (mirex)," 2013.

[92] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords." in *Proc. ISMIR*, 2010, pp. 135–140.

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

[94] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *Proc. ISMIR*, 2012, pp. 559–564.

[95] E. Humphrey, S. Durand, and B. McFee, "Openmic-2018: An open data-set for multiple instrument recognition." in *Proc. ISMIR*, 2018, pp. 438–444.

[96] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2016.

[97] O. Slizovskaia, E. Gómez Gutiérrez, and G. Haro Ortega, "Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies," in *Großmann R, Hajdu G, editors. Proceedings SMC 2016. 13th Sound and Music Computing Conference; 2016 Aug 31; Hamburg, Germany. Hamburg (Germany): ZM4, Hochschule für Musik und Theater Hamburg; 2016. p. 442-7.* Zentrum für Mikrotonale Musik und Multimediale Komposition (ZM4), Hochschule , 2016.

[98] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[99] P. Knees and K. Andersen, "Searching for audio by sketching mental images of sound: A brave new idea for audio retrieval in creative music production," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 95–102.

[100] M. Mauch, K. C. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription." in *Proc. ISMIR*, 2009, pp. 231–236.

[101] E. Cambouropoulos, "Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface," in *Joint International Conference on Cognitive and Systematic Musicology*. Springer, 1996, pp. 277–293.

[102] B. Janssen, P. Van Kranenburg, and A. Volk, "Finding occurrences of melodic segments in folk songs employing symbolic similarity measures," *Journal of New Music Research*, vol. 46, no. 2, pp. 118–134, 2017.

[103] J. B. L. Smith and M. Goto, "Nonnegative tensor factorization for source separation of loops in audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, 2018, pp. 171–175.

[104] D. Meredith, "Using siateccompress to discover repeated themes and sections in polyphonic music," *Music Information Retrieval Evaluation Exchange (MIREX)*, 2016.

[105] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models." in *Proc. ISMIR*, 2009, pp. 621–626.

[106] D. Herremans and E. Chew, "Tension ribbons: Quantifying and visualising tonal tension." in *Second International Conference on Technologies for Music Notation and Representation (TENOR)*, 2016.

[107] I. Barbancho, G. Tzanetakis, A. M. Barbancho, and L. J. Tardón, "Discrimination between ascending/descending pitch arpeggios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2194–2203, 2018.

[108] L. Yang, K. Z. Rajab, and E. Chew, "The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation," *Journal of Mathematics and Music*, vol. 11, no. 1, pp. 42–60, 2017.

[109] C. Baume, G. Fazekas, M. Barthet, D. Marston, and M. Sandler, "Selection of audio features for music emotion recognition using production music," in *Audio Engineering Society Conf.: 53rd Int. Conf.: Semantic Audio*. Audio Engineering Society, 2014.

[110] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proc. Int. Conf. on Multimedia information retrieval*. ACM, 2010, pp. 267–274.
[111] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.
[112] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *Proc. ICASSP*. IEEE, 2012, pp. 81–84.
[113] R. Panda, B. Rocha, and R. P. Paiva, "Music emotion recognition with standard and melodic audio features," *Applied Artificial Intelligence*, vol. 29, no. 4, pp. 313–334, 2015.
[114] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.
[115] M. Schedl, E. Gómez, J. Urbano *et al.*, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.
[116] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. ICASSP*. IEEE, 2018, pp. 161–165.
[117] J. Scott, E. M. Schmidt, M. Prockup, B. Morton, and Y. E. Kim, "Predicting Time-Varying Musical Emotion Distributions from Multi-Track Audio," *Proc. CMMR*, vol. 6, pp. 186–193, 2012.
[118] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research." in *Proc. ISMIR*, 2014, pp. 155–160.
[119] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2019.
[120] C. N. Reed and E. Chew, "Effects of musical stimulus habituation and musical training on felt tension," in *Late Breaking/Demo, Proc. ISMIR, Delft, The Netherlands*, 2019.
[121] C. Dromey, S. O. Holmes, J. A. Hopkin, and K. Tanner, "The effects of emotional expression on vibrato," *Journal of Voice*, vol. 29, no. 2, pp. 170–181, 2015.
[122] T. Eerola, A. Friberg, and R. Bresin, "Emotional expression in music: contribution, linearity, and additivity of primary musical cues," *Frontiers in psychology*, vol. 4, p. 487, 2013.
[123] P.-C. Li, L. Su, Y.-H. Yang, A. W. Su *et al.*, "Analysis of Expressive Musical Terms in Violin Using Score-Informed and Expression-Based Audio Features," in *Proc. ISMIR*, 2015, pp. 809–815.
[124] B. Liang, G. Fazekas, and M. Sandler, "Piano Sustain-pedal Detection Using Convolutional Neural Networks," in *Proc. ICASSP*. IEEE, 2019, pp. 241–245.
[125] L. Su, L.-F. Yu, and Y.-H. Yang, "Sparse cepstral, phase codes for guitar playing technique classification." in *Proc. ISMIR*, 2014, pp. 9–14.
[126] Y.-P. Chen, L. Su, Y.-H. Yang *et al.*, "Electric guitar playing technique detection in real-world recording based on f0 sequence pattern recognition." in *Proc. ISMIR*, 2015, pp. 708–714.
[127] C. Wang, E. Benetos, X. Lostanlen, and E. Chew, "Adaptive Time-Frequency Scattering for Periodic Modulation Recognition in Music Signals," in *Proc. ISMIR*, 2019.
[128] V. Lostanlen, J. Andén, and M. Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 2018, pp. 1–10.
[129] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 2, p. 47, 2019.
[130] E. Maestre, P. Papiotis, M. Marchini, Q. Llimona, O. Mayor, A. Pérez, and M. M. Wanderley, "Enriched multimodal representations of music performances: Online access and visualization," *IEEE Multimedia*, vol. 24, no. 1, pp. 24–34, 2017.
[131] C. C. Liem, E. Gómez, and G. Tzanetakis, "Multimedia technologies for enriched music performance, production, and consumption," *IEEE MultiMedia*, no. 1, pp. 20–23, 2017.

Simin Yang is a doctoral candidate at the Centre for Digital Music (C4DM) in the School of Electronic Engineering and Computer Science at Queen Mary University of London (QMUL). She received a Bachelor of Engineering in Electronic Information Engineering from Shandong University, China in 2015. Her research interests include emotion and sentiment analysis, human-computer interaction, human behaviour modelling and recommender systems.

Courtney N. Reed is a doctoral candidate at the Centre for Digital Music (C4DM) at Queen Mary University of London (QMUL). She received a BMus in Electronic Production and Design (Berklee College of Music, Boston, USA) in 2016 and an MSc in Computer Science (QMUL) in 2018. She is a member of the Augmented Instruments Lab, where she is focused on design for the voice and the vocalist's relationship with their instrument. Her research interests include musical gesture in vocal performance, intention and musical imagery use, vocal physiology measured through biosignals, namely surface electromyography, and embodied interaction design.

Elaine Chew is a senior Centre National de la Recherche Scientifique researcher in the Sciences et Technologies de la Musique et du Son Lab at the Institut de Recherche et Coordination Acoustique/Musique, and Visiting Professor of Engineering in the Faculty of Natural and Mathematical Sciences at King's College London. She is principal investigator of the European Research Council projects COSMOS (Computational Shaping and Modeling of Musical Structures) and HEART.FM (Maximizing the Therapeutic Potential of Music through Tailored Therapy with Physiological Feedback in Cardiovascular Disease). Her work has been recognised by an Presidential Early Career Award for Scientists and Engineers and an NSF Faculty Early Career Development award, and Fellowships at Harvard's Radcliffe Institute for Advanced Study. She is an alum (Fellow) of the National Academy of Sciences Kavli and National Academy of Engineering Frontiers of Science/Engineering Symposia. Her research focuses on the mathematical and computational modelling of musical structures in music and electrocardiographic sequences. Applications include modelling of music performance, AI music generation, music-heart-brain interactions, and computational arrhythmia research. As a pianist, she integrates her research into concert-conversations that showcase scientific visualisations and lab-grown compositions.

Mathieu Barthet is a senior lecturer in digital media at Queen Mary University of London (QMUL). He is a co-investigator of the UKRI EPSRC Centre for Doctoral Training in AI & Music for which he oversees industry partnerships, and programme coordinator of the MSc in Media and Arts Technology. He received an MSc degree in Electronics and Computer Science in 2003 (Paris VI University/Ecole Polytechnique de Montréal), and an MSc degree in Acoustics in 2004 (Aix-Marseille II University/Ecole Centrale Marseille). He was awarded a PhD in Acoustics, Signal Processing and Computer Science applied to Music from Aix-Marseille II University and CNRS-LMA in 2008, and joined the Centre for Digital Music at QMUL in 2009. He was co-investigator of the EU Audio Commons project and principal investigator of the EU MSCA-IF Internet of Musical Things and Innovate UK ALIVEmusic projects. He conducts multidisciplinary research on topics including music and emotions, musical timbre, intelligent musical interfaces, music recommendation, audiovisual interfaces and extended reality. He has served as General Chair of the Computer Music Modeling and Retrieval symposium on "Music and Emotions" (2012), and Program and Paper Chair of the ACM Audio Mostly conference on "Augmented and Participatory Sound and Music Experiences" (2017). His research on interactive music systems has lead to performances in venues such as Barbican Arts Centre, Wiltons Music Hall, Strasbourgs Cathedral, and CHI, ACII, CMMR, and BBC conferences.

APPENDIX A

A.1 Live Performance Segments

The recording of the performance used in this study can be found at: https://youtu.be/55JJLq3ewHs. The start and end times of the segments from the Live study as they occur in this video are as follows:

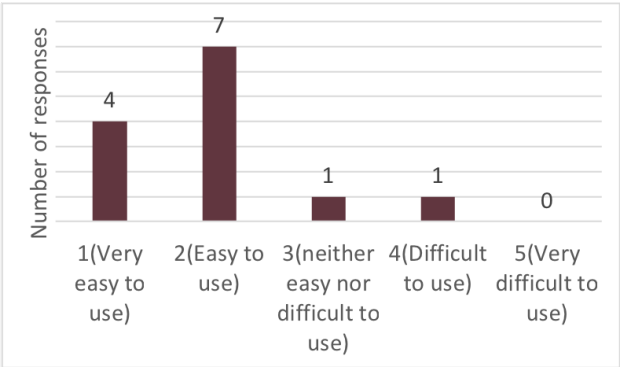
Movement	Segment	Time Start	Time End	Duration(sec)
1	1	00:00.0	01:03.5	63.5
	2	01:03.5	01:56.3	52.9
	3	01:56.3	02:30.8	34.4
	4	02:30.8	03:03.9	33.1
	5	03:03.9	03:40.5	36.6
	6	03:40.5	04:06.0	25.5
	7	04:06.0	05:05.8	59.8
	8	05:05.8	05:42.4	36.6
	9	05:42.4	06:05.5	23.2
	10	06:05.5	06:36.5	31.0
	11	06:36.5	07:50.7	74.1
	12	07:50.7	08:37.6	46.9
	13	08:37.6	09:15.6	38.0
	14	09:15.6	09:54.3	38.7
	15	09:54.3	10:14.6	20.3
	16	10:14.6	11:13.7	59.2
2	17	11:13.7	12:02.1	48.4
	18	12:02.1	12:29.5	27.3
	19	12:29.5	13:00.2	30.8
	20	13:00.2	13:27.7	27.5
	21	13:27.7	13:53.8	26.1
	22	13:53.8	14:40.6	46.7
	23	14:40.6	15:16.1	35.5
	24	15:16.1	16:27.0	70.9
	25	16:27.0	17:01.0	41.5
3	26	17:08.6	17:25.0	16.4
	27	17:25.0	17:43.2	18.2
	28	17:43.2	18:00.1	16.9
	29	18:00.1	18:18.1	18.1
	30	18:18.1	18:35.8	17.6
	31	18:35.8	18:50.8	15.1
	32	18:50.8	19:12.1	21.2
	33	19:12.1	19:40.3	28.2
	34	19:40.3	19:57.0	16.7
	35	19:57.0	20:07.2	10.2
	36	20:07.2	20:19.9	12.7
	37	20:19.9	20:31.1	11.2
	38	20:31.1	20:46.6	15.5
	39	20:46.6	21:17.2	30.6
	40	21:17.2	21:41.1	23.9
	41	21:41.1	21:56.5	15.4
	42	21:56.5	22:17.8	21.4
	43	22:17.8	22:51.7	33.9
	44	22:51.7	23:25.5	33.7
	45	23:25.5	23:52.9	27.5

Tbl. 8: The 45 rehearsal segments in the Live recording of Arno Babajanian's *Piano Trio in F # minor*, with corresponding beginning and end times and duration.

A.2 Live Study Questionnaire Responses

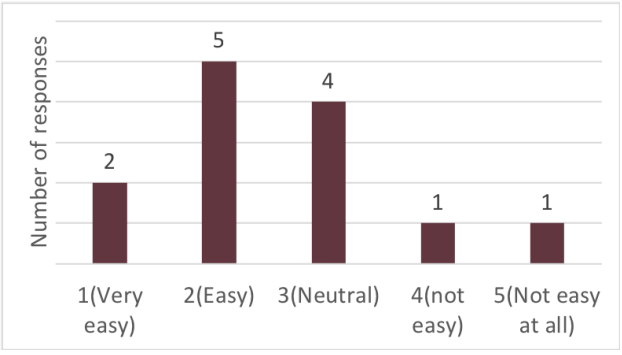
The questions and responses for the Mood Rater evaluation questionnaire are summarised as follows:

Question A. 11 out of 13 participants indicated that the Mood Rater app was easy to use. From the corresponding comments, participants evaluated the Mood Rater app as presenting a "*high degree of dexterity*," and assessed it as being "*simple to view and touch.*" Some participants evaluated the two-dimensional valence and arousal plane as "*easy to understand*," "*easy to use*," and that the colour chart helped to understand the VA dimensions. One participant who rated Mood Rater as "*difficult to use*" explained why: "*During the live music, I have to pay attention to the concert; using the app will distract me from understanding the music. Could you change it to testing the pulse of fingers to identify the emotion?*" This indicates that using a smartphone app for live annotation during a concert might not be ideal for every participant. Biosensors could be useful to monitor emotions without requiring direct input from participants: however, they pose other challenges, such as data noise and ethical considerations.



Q-A: "*Overall, how easy to use do you find Mood Rater?*"

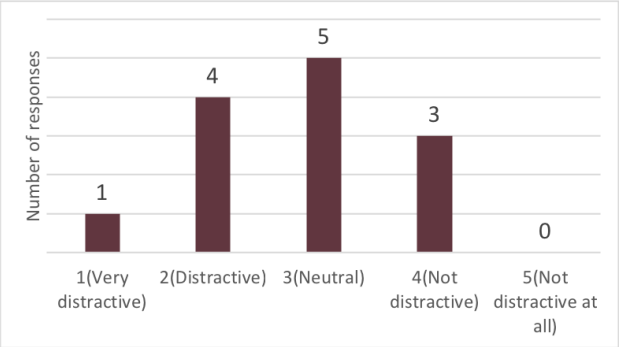
Question B. For the second question, 7 of 13 participants evaluated the task of rating perceived emotions during the performance as *Easy*; four participants choose *Neutral*; two participants evaluated the task as not easy to complete. Some participants expressed the following: "*The music is kind of not that easy to get the exact emotion*," "*sometimes the emotion changes too fast to track*," "*Some emotion is very difficult to explain.*" These comments may indicate that self-reporting perceived emotions while listening to music can be intricate due to the style or complexity of the music, the rapidity of change, or the difficulty in identifying the emotions using the proposed model (in our case valence and arousal).



Q-B: "*How would you judge the degree to which using the Mood rater app distracted you from the performance?*"

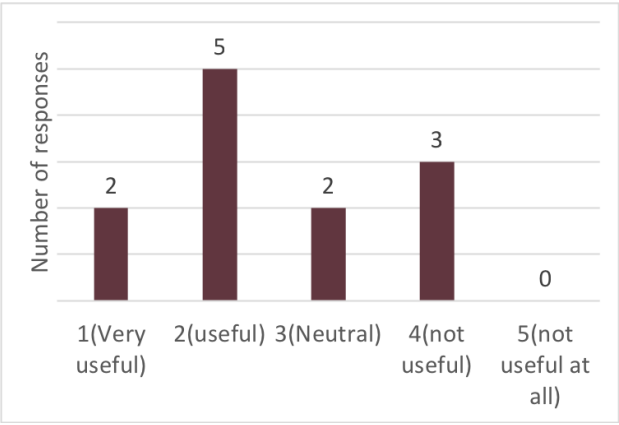
Question C. 3 out of 13 participants commented that the rating process did not distract them from the performance (*Not distractive*), 5 participants chose *Neutral*; 5 indicated that the rating process was *Distractive* from the performance. Reasons included: "*I have to watch this app except for the vision of the stage*," "*looking at my phone and deciding where my mood related to the axis took a little away from experience*," "*It was not*

distractive for me, but I was curious it could distract the others. An interface with a less bright colour would be better." These reveal some limitations of using smartphones to annotate emotion in live music.



Q-C: *"How would you judge the degree to which using the Mood rater app distracted you from the performance?"*

Question D. 3 out of 13 participants indicated the tags were *not useful*, 2 remained *neutral*, and the remaining 8 participants evaluated the mood tags as *useful*. People who held an unfavourable view found the tags to be inaccurate and not adapted to the music, or felt that sometimes they did not match their current emotional state.



Q-D: *"Mood Rater app shows some words (so called mood tags) when a specific part of the screen is selected. Were these useful?"*

A.3 Live Study Open-Ended Responses

The results of the open-ended question to improve the Mood Rater app are summarised as follows:

Mood Rater Tags: Mood tags placed underneath the interface can sometimes be confusing. The tags should be revised or removed (the latter case is not ideal since several participants mentioned that the tags helped them); a participant suggested an extra button be added to let users choose if the tag is correct or not.

Mood Rater Interface: The size of the Mood Rater interface does not fit all types of smartphone screens, which makes the application sometimes less aesthetically pleasing and engaging. Some participants suggested an interface with higher resolution and a slightly darker colour. One participant suggested that some facial expression icons, for instance emojis, could be added on the VA plane to help the understanding of the two dimensions.

Extra Functions: One participant suggested adding a counting function to count how many times the participant has tapped on the screen, to make the app more engaging. Another suggested adding a function whereby the user could put their finger on the flash and camera to monitor the user's heartbeat and get extra physiological information during the rating process. One participant suggested that Mood Rater should prevent or suppress notifications from other apps from appearing, which potentially interrupts the rating process.