MULTI-SCALE ANALYSIS OF AGREEMENT LEVELS IN PERCEIVED EMOTION RATINGS DURING LIVE PERFORMANCE

Simin Yang, Mathieu Barthet, Elaine Chew
Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London
{simin.yang, m.barthet, elaine.chew}@qmul.ac.uk

ABSTRACT

This study characterises and analyses the dynamics of agreement on perceived emotions in the context of live music performance. We record emotion annotations from 15 participants. The recorded information is analysed at multiple timescales: whole piece, entire movements, sections defined by rehearsal numbers, and individual bars. Inter-rater reliability (IRR) is estimated using the intra-class correlation coefficient (ICC) at each time scale. Strong agreement is found on perceived arousal and valence at the whole piece and movement time scales, with stronger agreement for arousal than valence. At the rehearsal segment level, strong agreement is found in 17/45 segments for arousal and 7/45 for valence. Agreement was extremely low, mostly near zero, at the bar level. Finally, we posit that agreement is a mark of emotion clarity and disagreement an indicator of emotion complexity.

1. INTRODUCTION

The issue of subjectivity in perceiving emotion in music has been the topic of numerous studies [1]. Considerably less attention has been paid to understanding the relationship between musical attributes and listeners' agreement on perceived emotion. Such a study on emotional response may help identify musical segments representing moments of emotion "clarity" or lack there of, which can benefit semantic segmentation and music summarisation. As a first step, we investigate the issue of subjectivity in emotion perception by examining raters' agreement across time-based emotion ratings made by 15 participants during a live performance of a contemporary three-movement chamber music piece.

2. MEASURING RATING CONSISTENCY

Inter-rater reliability (IRR) measures the degree of agreement in independent ratings from two or more raters [3]. The intra-class correlation (ICC) is commonly employed when assessing IRR for ordinal, interval, or ratio variables [4]. Higher ICC values correspond to higher degrees of IRR; an ICC value of 1 indicates total agreement, while an ICC value of 0 represents random agreement. Negative ICC values are also possible, indicating systematic disagreement.

3. EXPERIMENT DESIGN

We set up a study to collect listeners' time-varying emotion perceptions in a live music performance held at Queen Mary University of London on 22 October 2015 as part of the Inside Out Festival.

The music stimuli consist of a performance of the three movements of the *Piano Trio in F# minor by Arno Babajanian (1921 - 1983)*. The movements are of widely disparate characters: the first is marked *Largo—Allegro expressivo—Maestoso*, indicating its predominantly slow and noble tempo with a faster middle part; the second is marked *Andante*, at a walking pace; and, the third is marked *Allegro vivace*, which is lively and rapid. The entire performance is about 23 minutes in length. We use the rehearsal marks, created to facilitate rehearsing, as a guideline to partition the music into 45 segments. There are 16, 9 and 20 segments in movements 1, 2, and 3, respectively, each lasting between 11 and 72 seconds.

For data collection, we used a real-time interactive tool, the "Mood Rater" [1], shown on the left side of Figure 2. The "Mood Rater" is derived from the "Mood Conductor" [2], designed originally for use in participatory performance. The tool is a smartphone-friendly web application with an interface based on the Arousal-Valence (AV) space.

Before the concert, audience members are instructed on how to access the "Mood Rater" app using their mobile devices. They are also given information on the nature of the study and the AV representation. Finally, they were invited to annotate their perceived emotion in the AV space during the performance. No specific indications are given to

[1] http://bit.ly/moodxp2

participants as to how often they should give ratings. We thus assume that participants annotate their perceived emotions whenever they detect a change, and that the previous annotation persists until a different emotion is reported.

4. PRELIMINARY RESULTS

Over the course of the whole piece (23 minutes), a total of 1023 emotion annotations are collected from the 15 participants. Figure 1 shows how the audience rating frequency varied over the course of the performance and from one participant to another. Right side of Figure 2 gives the distribution of the 1023 AV ratings. The collected data spans all four quadrants of the AV space, whereas less ratings in the lower left quadrant, which reflects the wide variety of expressions within the piece. This summary of the emotion ratings also suggests that a single emotion rating would not suffice for characterising the whole piece.

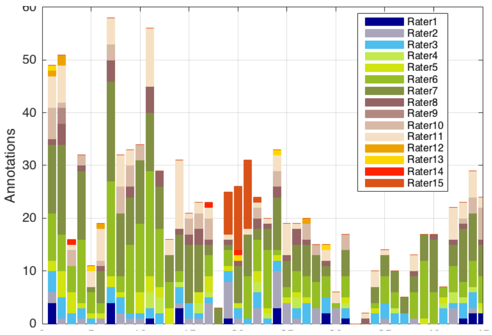

Figure 1. Distribution of annotations over reh segments.

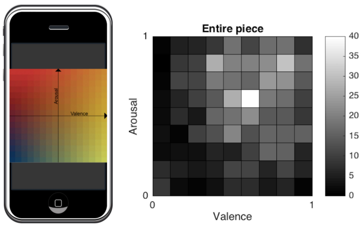

Figure 2. The "Mood Rater" data collection tool (left) and distribution of VA ratings over the whole piece (right)

To quantify consistency in participant ratings, intra-class correlation (ICC) was computed at all time scales: whole piece, entire movement, rehearsal segment, and bar. Individual AV ratings were re-sampled using a sampling period of 1 second, and the results are reported in Figure 3.

Strongly significant agreement is found for arousal at the whole piece level and in all three movements, whereas significant to strongly significant agreement, even though much lower, is also found for the valence ratings. The ICC at the rehearsal segment level varies a great deal over time (SD=0.27 (Arsoual), SD=0.15 (Valence). Segments with disagreement emerged at the rehearsal segment time-scale in both arousal and valence ratings, with some segments presenting strong agreement (p-value $<$ 0.05 in 17 segments out of a total of 45 for arousal and 7 out of 45 for valence) and others disagreement (having negative ICC in 11 segments for arousal and 17 for valence). We propose that segments with strong agreement project a clear emotion, while those showing disagreement present more complex emotions subject to variable interpretation. At the bar level, the agreement was extremely low, all near zero except for a few low ICC numbers around 0.2, these results show that the experiment listeners did not express change of perceived emotions at the same time, however tended to agree when looking at longer time scales.

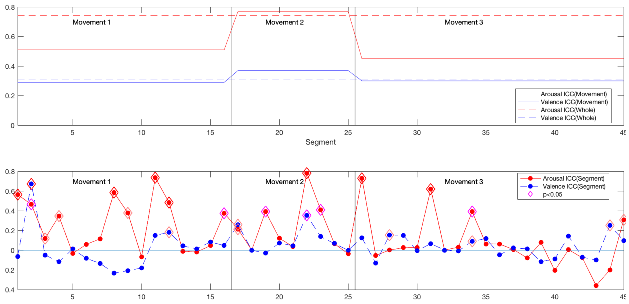

Figure 3. Intra-class correlation (ICC) for arousal and valence ratings for whole piece, movement, and rehearsal segment.

Overall, participants agreed on perceived arousal and valence at the whole piece and movement time scale, with arousal showing stronger agreement than valence. At the rehearsal segment level, participants still tend to agree more on arousal than valence in most segments. This is consistent with prior music emotion recognition studies, which showed valence to be more difficult to predict.

5. ACKNOWLEDGMENTS

This work was funded, in part, by the China Scholarship Council and the EU H2020 Audio Commons grant (No.688382).

6. REFERENCES

[1] Tuomas Eerola. Modeling emotions in music: Advances in conceptual, contextual and validity issues. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[2] Gyorgy Fazekas, Mathieu Barthet, and Mark B Sandler. The mood conductor system: Audience and performer interaction using mobile technology and emotion cues. In *10th International Symposium on Computer Music Multidisciplinary Research (CMMR13)*, pages 15–18, 2013.

[3] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[4] Kevin A Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.