Towards an experiment on perception of affective music generation using MetaCompose

Marco Scirea
University of Southern Denmark

Peter Eklund
Deakin University, Australia

Julian Togelius
New York University, USA

Sebastian Risi
IT University of Copenhagen, Denmark

ABSTRACT

MetaCompose is a music generator based on a hybrid evolutionary technique combining FI-2POP and multi-objective optimization. In this paper we employ the MetaCompose music generator to create music in real-time that expresses different mood-states in a game-playing environment (Checkers) and present preliminary results of an experiment focusing on determining (i) if differences in player experience can be observed when using affective-dynamic music compared to static music; and (ii) if any difference is observed when the music supports the game's internal narrative/state. Participants were tasked to play two games of Checkers while listening to two (out of three) different set-ups of game-related generated music. The possible set-ups were: static expression, consistent affective expression, and random affective expression.

CCS CONCEPTS

• Theory of computation → Evolutionary algorithms; **• Applied computing → sound and music computing**;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Marco Scirea, Peter Eklund, Julian Togelius, and Sebastian Risi. 2018. Towards an experiment on perception of affective music generation using MetaCompose . In *Proceedings of the Genetic and Evolutionary Computation Conference 2018 (GECCO '18 Companion)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3205651.3205745

1 INTRODUCTION

This paper reports the use of the MetaCompose [3, 4] music generator in a game-playing context. We previously showed that the MetaCompose system was able to reliably express moods in music [2], however we have not to this point demonstrated that affective music system can improve player experience. In this paper we aim to provide a first look at how much (if any) effect the use of this kind of music generator has on the perception of a game. The game chosen for this experiment is American-checkers. Checkers is chosen for three reasons: it is well-known, has simple rules easily understood even for players that are unfamiliar, and it has a minimal amount of intrinsic narrative. While the two former choices derive from practical considerations, we also wanted a game that satisfied the latter requirement to remove as many variables as possible that could influence perceptions of the game. The research questions this study addresses are:

(1) can any difference in player experience (emotionally) be observed when presented with affective-dynamic music compared to static music?
(2) can differences be observed when the music supports the game's internal narrative/state?

In short, the first allows us to explore the effect of *dynamic* music, while the second tests the effect of *adaptive* music. To this end we present and discuss the results of a participant-based evaluation in which test-subjects are tasked to play two Checkers games with musical accompaniment. Both self-reported data is collected – through a questionnaire – as well as physiological data via sensor input.

The results support the two hypothesis, especially in showing preference towards dynamic affective music and showcase the potential of the MetaCompose system – and by extension other affective music generators. More importantly, they suggest that the affective/dynamic music generation paradigm on which MetaCompose is based can lead to an improved player experience.

2 METACOMPOSE

MetaCompose [3, 4] consists of three main components: (i) *composition generator*, (ii) *real-time affective music composer*. This section presents a summary of the music generation method employed by MetaCompose, a more complete description can be found in [4]. The *composition generator* (i) creates the basic abstraction of a score used by the *real-time affective music composer* in order to (ii) generate the final score according to a specific mood or affective state. In other words, the *composition generator* (i) serves as a composer that only writes the basic outline of a piece, while the *real-time affective music composer* (ii) acts as an ensemble, free to interpret the piece in different ways. The system also has an *archive* which maintains a database of all the previous compositions connected to the respective levels/scenes of the game-state while also allowing a rank to be computed that measures the novelty of future compositions compared to those previously generated. MetaCompose is designed to react to game events depending on the effect desired. Examples of responses to such events include: a simple change in the affective state, a variation of the current composition, or an entirely new composition.

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07…$15.00
https://doi.org/10.1145/3205651.3205745

3 EXPERIMENT DESIGN

An experiment was devised where participants would play two games of checkers, while listening to two (out of three) different set-ups of accompanying music. The three possible experimental set-ups were: static expression, consistent affective expression, and random affective expression. Afterwards, participants were tasked with answering four comparative questions regarding the two games. Martinez and Yannakakis [1] suggest that ranking produces more consistent and reliable data when annotating affect information, participants are therefore asked to compare two pieces of music and rank them. The questions are:

- **Which game did you find more engaging?**
  "The first one"/"The second one"/"Neither"/ "Both Equally"
- **In which game was the music best?**
  "The first one"/"The second one"/"Neither"/ "Both Equally"
- **In which game did the music better match how exciting the game was?**
  "The first one"/"The second one"/"Neither"/ "Both Equally"
- **In which game did the music better match how well you were playing?**
  "The first one"/"The second one"/"Neither"/ "Both Equally"

Also included were the more neutral answers "Neither" and "Both Equally" to avoid randomness in the data from participants who cannot decide which clip satisfies the evaluation criterion. A survey was prepared with HTML and PHP, using a MySQL database to hold the data collected. The Checkers-framework used is an open-source AI framework called *raven-checkers*[1] which includes provision of a computer game-playing agent. The PHP code externally invokes the Checkers-framework through the *exec()* function, which effectively stops the execution of the PHP code until the game terminates. The experiment was designed for the participants to play two games of Checkers with 2 randomly chosen set-ups (repetitions of the same set-up were not allowed). As each game can take between 5 and 10 minutes, the experiment was designed to last between 10 and 20 minutes for each participant.

The music in this experiment is not generated beforehand, instead the score is generated in real-time by MetaCompose. For each participant, MetaCompose creates a single composition and uses that as the basis for the music generated in both of the player's play-through of both Checkers set-ups. In this way we ensure no difference in the player's response due to a potential quality difference between two compositions, the baseline accompaniment is identical.

4 RESULTS AND ANALYSIS

The data collected corresponds to 29 (19 males, 9 females, and 1 participant did not express gender) self-reported comparisons. The participants' age has an average of ≈ 28.9 years ($stdev \approx 5.9$). In regards to the other demographic features, expressed in 5-point Likert scale (0–4), most people self-reported little prior experience with the game of Checkers ($avg = 0.89$, $stdev \approx 0.87$, $mode = 1$), and a considerable experience with computer videogames ($avg = 2.68$, $stdev \approx 1.05$, $mode = 2$). No matter how we divide the population, the results are not significantly different,

[1]https://github.com/bcorfman/raven-checkers

Table 1: The differences between set-ups shown as p-values calculated using a two-tailed binomial test. Statistically significant values are in bold.

	Engage	*Best*	*Exciting*	*Well*
Consistent/random	5.47E-02	**7.81E-03**	7.03E-02	2.19E-01
Consistent/static	**1.76E-02**	**3.13E-02**	5.47E-02	2.50E-01
Random/Static	1.64E-01	2.34E-01	1.09E-01	5.00E-01

possibly because of the limited number of participants and their relative homogeneity. As shorthand we will refer as the criteria using the labels: *engage*, *best*, *exciting*, *well*. Refer to Section 3 for the complete text of the questions.

In this work, only definitive answers are considered (i.e. the participant chooses one of the music clips presented). Under the definite choice constraint, the data therefore becomes Boolean: the answers are either *"user preferred the first set-up"* or *"user preferred the second set-up"*. To analyse such data a two-tailed binomial test is used, with as null hypothesis that both categories are equally likely to occur and, as we have only two possible outcomes, that probability is 0.5. The values calculated are shown in Table 1.

5 DISCUSSION

The self-reporting task of the experiment showed how music with 'affect expression' – consistent with the game-state – appears better perceived than the other two set-ups. In particular, we observe it is perceived as: 'having better overall quality', 'leading to a more engaging experience', and – to a lesser degree – 'better matching the perceived excitement in the game'. The static set-up and the random expression set-up appear to be more equivalent, although we can observe some non-significant differences between the two: the static set-up seems to be generally better perceived ('more engaging' and 'overall better quality'), while the random set-up seems to better match the perceived in terms of 'excitement of the game'. We hypothesize that the random-group has too many disruptive changes in expression to be particularly well-liked by the listener, while the static-group by definition does not present any changes that might match game-play, leading to lower ratings in regards to the excitement criterion. When looking at the answers for the last criterion ("in which game did the music better match how well you were playing?") we find inconsistent results between each of the groups. This may be caused by the complexity of the question which requires the participant also to evaluate his/her own game-play performance compared with performance in the second game.

REFERENCES

[1] Hector P Martinez, Georgios N Yannakakis, and John Hallam. 2014. Don't classify ratings of affect; rank them! *IEEE Transactions on Affective Computing* 5, 3 (2014), 314–326.

[2] Marco Scirea, Peter Eklund, Julian Togelius, and Sebastian Risi. 2017. Can you feel it?: evaluation of affective expression in music generated by MetaCompose. In *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 211–218.

[3] Marco Scirea, Julian Togelius, Peter Eklund, and Sebastian Risi. 2016. MetaCompose: A Compositional Evolutionary Music Composer. In *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, 202–217.

[4] Marco Scirea, Julian Togelius, Peter Eklund, and Sebastian Risi. 2017. Affective evolutionary music composition with MetaCompose. *Genetic Programming and Evolvable Machines* 18, 4 (2017), 433–465.