# Emotions in AI-generated music compared to emotions in human-composed music

Maria Polak

Aalborg University Copenhagen, Sound and Music Computing, 7th Semester

## Abstract

This paper presents a pilot experiment which compares emotions induced through music composed by two different types of composers, namely AI and human. The experiment was conducted in a form of an online survey and the participants were recruited through social-media. To measure the strength of the induced emotions, Likert scales, derived from the three-dimensional model, were used. The results were later analysed with a Wilcoxon signed-rank test. Early insights suggest that emotional response to AI-composed music is strong, but different from traditional human-composed music.

## 1 Introduction

With the latest growth of generative AI we might wonder if AI-composed music should be considered music. While this discussion was the main inspiration, the purpose of this study was to assess inseparable part of music, emotions, in AI-generated music compared to human-composed music.

### 1.1 Hypothesis

*AI-generated music can induce emotions to similar extent as human-composed music and the type of composer does not matter.*

## 2 Methods

### 2.1 Model

The idea behind the experiment was to measure emotions induced by AI and human-composed music and assess if there are obvious differences in emotional response between the types of composers. In other words, the main interest of this study was comparison of the induced emotions. Because of that, the most important factor for choosing the rating model was its stability and reliability. In the end, model created by Ulrich Schimmack and Alexander Grob [8] was chosen, as it produced reliable results when rating felt [10] and perceived [4] emotions in previous studies.

### 2.2 Participants

Participants were recruited through social-media and internal communication means of Aalborg University. In total, 18 people completed the questionnaire. During experiment, participants were asked to provide standard demographic information, as it can be used as moderating variable and can create additional insight into data [3]. The participants were of varying age (50% 25-34, 16.7% 35-44, 33.3% 45-54) and gender (72.3% Male, 27.7% Female). Additionally, participants varied in their background musical training (61.1% no musical training, 22.2% with formal music education, 5.6% with informal music education and 11.1% other) and number of hours they spend listening to music daily (5.6% does not listen to music daily, 11.1% up to 30 min, 27.8% 30 min to 1 hour, 44.4% 1 to 3 hours, 11.1% more than 3 hours).

### 2.3 Stimuli

In total four excerpts were selected as stimuli: two for each composer type. All excerpts belonged to the same genre - cinematic orchestral music. This genre was selected as the most important task of cinematic music is,

in general, affecting listener's emotions.

Each human-composed excerpt was paired with one of the AI-generated excerpts in terms of the main communicated emotion cue. First pair represented sad and relaxing emotions, while the other represented tenderness. For human-composed excerpts, two pieces, *Beyond Good and Evil* and *I will find you*, created by a music production company *Audiomachine* were used.

The AI-composed pieces were generated using state-of-the-art music generation AI, AIVA [1]. AIVA is based on stochastic algorithms. It offers 5 ways to generate pieces – based on music style, chord progression, step-by-step (user specifies the style, chord progressions and instruments), from an influence (based on another music piece) and predefined presets (legacy). Multiple excerpts were generated using these methods, and in the end two excerpts were selected. As previously stated, both excerpts were chosen to match emotional cues from human-composed versions.

Authors of [5] suggested that 30 to 60 seconds long excerpt is needed to induce emotions. To fulfill that requirement, the selected excerpts were on average 70 second long, ranging between 60 and 90 seconds.

### 2.4 Procedure

The listening experiment was conducted in a form of an online survey on the Go Listen platform introduced and developed by the authors of [2].

The participants were informed that during the experiment they will rate emotions felt while listening to four music excerpts, and that the survey will take approximately 11 min to complete. The definitions of felt emotions, as well as difference between felt and perceived emotions, were explained to them before the core part of the test. Moreover, between listening to the excerpt and rating the emotions, the participants were reminded that they are asked to rate felt emotions and that it is okay to not feel any emotions.

After each excerpt, a block of 7 questions appeared on participant's screen. Each block consisted of 6 scales derived from the three-dimensional model (Pleasure, Displeasure, Awareness, Sleepiness, Tension, Relaxation) and a question ensuring that the participant did not recognize excerpt's source piece. Each emotion's strength was rated on a 7-point Likert scale where 1 stands for "No emotion felt" and 7 stands for "Very strong". To reduce bias, the listening test was available to users in two variants, each with a different order of the music excerpts.

To assure consistency of the testing environment, each participant was asked to wear headphones, if possible. In addition, a question about noisiness of their surroundings and volume calibration with a sample sound were part of the test. Until the end of the survey, participants were not informed that half of the excerpts were generated by an AI, since research suggest bias against AI-generated music [9] and other forms of AI-generated art [7].

## 3 Results

The experiment followed a within-subject design and its results are non-parametric ordinal data. Taking that into account, Wilcoxon signed-rank test [11] was chosen as a data analysis method. Within Wilcoxon method there are different approaches for handling pairs of observations with equal values (zeros). Since zeros often appear in discrete data, a Pratt [6] method for resolving them was used. The following hypotheses have been established:

- **Null hypothesis ($H_0$):** There is no difference in the induced emotions' ratings by music among AI composer and human composer across tested emotions.
- **Alternative hypothesis ($H_a$):** There is a difference in the induced emotions' ratings by music among AI composer and human composer across tested emotions.

As this is an exploratory study with a low sample size, a higher significance level for p-values was used - 0.1 (whereas a commonly used value is 0.05). The pairs of related samples were defined as follows, each emotion was compared between composers in three possible variants: relaxed AI - relaxed human, tender AI - tender human, all AI - all human.

Two of the participants recognized tender human-composed excerpt, since this could produce highly unreliable results, with potentially stronger emotional response, their responses were omitted during data analysis.

The calculated p-values greatly varied in values (Table 1), with the 5 lowest meeting the significance level criteria ($< 0.1$), whereas highest reaching value of 0.917. The findings suggest that null hypothesis is rejected, and that there is a significant difference in emotions felt when listening to music composed by different composers. This can also be seen and, to some extent, deducted from the results' visualisation (Fig 1). As seen in the plots, emotion ratings in AI-composed music were notably less consistent than the ratings in human-composed pieces, and in a few cases spanned a different range of values.

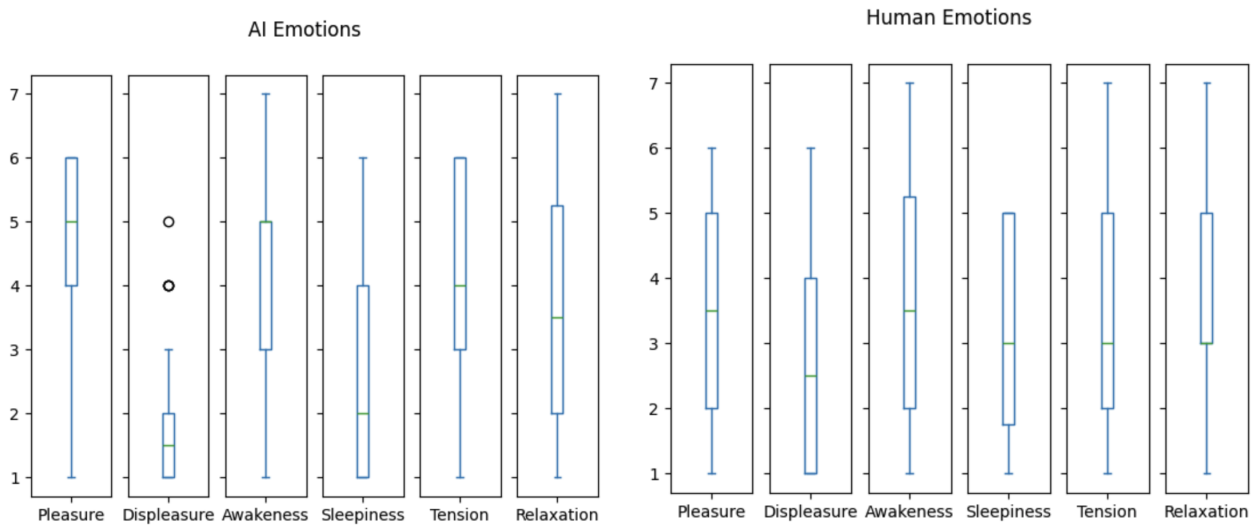


Figure 1: Emotion ratings depending on the type of composers

What is worth noting is the high gap between the ratings of pleasure and displeasure in AI generated music. One of the possible reasons for that could be the design of AI's generation mechanism. It is highly plausible that the authors of the AI wanted it to generate music which is considered pleasing for the users.

# 4 Discussion

In this part, the experiment is reviewed and potential improvements for future work are suggested.

First and foremost, excerpts used in a experiment could be determined by a pilot study. This would ensure similar characteristics of each pair AI and human composed music and possibly render the results more consistent

What is more, after completing the survey few participants said that they were surprised by half the excerpts being generated by an AI composer, that could suggest that the inconsistencies in the measured emotion were not dictated by the type of composer, but by the different characteristics of the excerpts.

In addition, some of the participants, especially the ones recruited from Aalborg University, were aware of the main goal of the experiment before completing it themselves, which might have introduced some unwanted bias in the resulting data.

Finally, the number of participants was not high enough, compromising the statistical reliability of the data. Even though the group was fairly diverse in ages and musical education background, this solely cannot overcome small amount of responses registered.

# 5 Conclusion

To conclude, findings of this research suggest that there is a difference in emotions induced by AI generated music in comparison to human-composed music. This does not suggest, that the emotional response is worse or weaker; the participants registered broad range of various emotions felt during listening of the AI composed music. Considering future work, the main take from this exploratory study is to consider a different, more structured approach to AI generated excerpt's selection.

| | AI vs Human | | Tender AI vs Tender Human | | Relaxed AI vs Relaxed Human | |
|---|---|---|---|---|---|---|
| | p-value | Effect Size | p-value | Effect Size | p-value | Effect Size |
| Pleasure | 0.053 | 0.29 | 0.093 | 0.25 | 0.263 | 0.33 |
| Displeasure | 0.021 | 0.23 | 0.165 | 0.28 | 0.072 | 0.19 |
| Awareness | 0.249 | 0.35 | 0.390 | 0.37 | 0.483 | 0.32 |
| Sleepiness | 0.191 | 0.34 | 0.531 | 0.38 | 0.180 | 0.27 |
| Tension | 0.251 | 0.33 | 0.916 | 0.44 | 0.057 | 0.16 |
| Relaxation | 0.791 | 0.46 | 0.917 | 0.47 | 0.835 | 0.46 |

Table 1: p-values and effect sizes across different comparisons

# References


[1] AIVA AI. https://creators.aiva.ai/. Accessed: 2023-11-22.

[2] Dan Barry, Qijian Zhang, Pheobe Wenyi Sun, and Andrew Hines. Go listen: An end-to-end online listening test platform. *Journal of Open Research Software*, 2021.

[3] John W. Creswell and J. David Creswell. *Research design : qualitative, quantitative, and mixed methods approaches*. SAGE, Thousand Oaks, California, sixth edition, international student edition. edition, 2023.

[4] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[5] Tuomas Eerola and Jonna K. Vuoskoski. A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307–340, 2013.

[6] John W. Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667, 1959.

[7] Martin Ragot, Nicolas Martin, and Salomé Cojean. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? CHI EA '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Ulrich Schimmack and Alexander Grob. Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.

[9] Daniel B Shank, Courtney Stefanik, Cassidy Stuhlsatz, Kaelyn Kacirek, and Amy M Belfi. Ai composer bias: Listeners like music less when they think it was composed by an ai. *Journal of experimental psychology. Applied*, 29(3):676–692, 2023.

[10] Jonna K. Vuoskoski and Tuomas Eerola. Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, 15(2):159–173, 2011.

[11] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.


# A Source Code and Audio Files

For the source code of all the code and audio files visit:
https://github.com/mp-smc23/MPC-Mini-Project

# B Experiment Instructions

Welcome!
In this experiment, you'll be listening to **4 music excerpts** and sharing your emotional responses. Your task is to rate the emotions you feel while listening to each excerpt. Additionally, you will be asked a few questions about your music preferences and background. The survey takes **approx. 11 minutes** to complete.
It is recommended you wear headphones. If you don't have any, you can use speakers instead.
This survey is being conducted as part of an academic study focusing on emotions in music for Sound and Music Computing Programme at Aalborg University. If you have any questions about this survey or the research, please do not hesitate to contact me at *mpolak23@student.aau.dk*

In this test you will always be asked about the **felt emotions**.
**Felt emotions refer to the actual emotions or feelings you experienced while listening to the music excerpt**. It is about how the music made you feel, the emotions it evoked, and your emotional state during or after listening. **Felt emotions** are personal and based on your individual experience. They are DIFFERENT from *perceived emotions*, which are the emotions you think the music tries to communicate or induce.
Ex. You can listen to a very happy song (*perceived emotion*), but be annoyed by it (**felt emotion**)

Whole survey is available at https://golisten.ucd.ie/task/audio-labeling/657f62dcf1a04554f996be67