Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Humanities
Department of Music

Aleksander Johnsen Solberg

# Affective music

## Using AI to generate music with emotion

Bachelor's thesis in Music Technology
Supervisor: Daniel Buner Formo
May 2022

Aleksander Johnsen Solberg

# Affective music

## Using AI to generate music with emotion

Bachelor's thesis in Music Technology
Supervisor: Daniel Buner Formo
May 2022

Norwegian University of Science and Technology
Faculty of Humanities
Department of Music

NTNU
Norwegian University of
Science and Technology

# Abstract

This bachelor's thesis addresses the task of generating music which conveys specific emotions through its musical expression using Artificial Intelligence. The project involves creating a new dataset consisting of 500 MIDI-files, together with information about their emotional expression in the form of valence and arousal. The files are divided into four quadrants based on this data, and is then used to train a musical generative algorithm in order to achieve four different models that are able to generate new music with the corresponding emotional expressions.

The MIDI-files used are sourced from the existing dataset ADL Piano MIDI, and information about valence and arousal is obtained through annotation done by human participants.

In the result, we see that this approach can have great potential, but that the generated music is very varied, primarily in the musical quality overall, but also in clarity of which emotion is conveyed. However, we see tendencies of the algorithm being able to generate music with the specified feeling.

# Sammendrag

Denne bacheloroppgaven tar for seg oppgaven med å generere musikk som formidler spesifikke emosjoner gjennom dens musikalske uttrykk ved bruk av Kunstig Intelligens. Prosjektet går ut på å lage et nytt datasett som består av 500 MIDI-filer, sammen med informasjon om deres følelsesmessige uttrykk i form av valens og energi. Filene er delt opp i fire kvadranter basert på disse dataene, og er så brukt til å trene opp en musikalsk generativ algoritme for å oppnå fire ulike modeller som kan generere ny musikk med de fire korresponderende følelsesuttrykkene.

MIDI-filene er hentet fra det eksisterende datasettet ADL Piano MIDI, og informasjon om valens og energi er skaffet gjennom annotering gjort av menneskelige deltagere.

I resultatet ser vi at denne fremgangsmåten kan ha stort potensiale, men at den genererte musikken er svært variert, hovedsakelig i den musikalske kvaliteten totalt sett, men også i tydeligheten av hvilken emosjon som er formidlet. Vi ser derimot tendenser til at algoritmen klarer å generere musikk med den spesifiserte følelsen.

## Acknowledgements

I would like to thank Agnete Djupvik for the idea of this project, and for help and guidance. I would also like to thank Martin Jensen for his very generous help in making the survey. Finally, I would like to thank my supervisor, Daniel Formo, for his help and guidance.

# Contents

# Acronyms

**AI** Artificial Intelligence. 1, 3, 5–8, 10, 13, 18

**AMC** Affective Music Composition. 5–8, 18

**MASESTRO** MIDI and Audio Edited for Synchronous TRacks and Organization. 9, 11

**MCC** Musical Computational Creativity. 6, 7, 15

**MER** Music Emotion Recognition. 7–9

**MIDI** Musical Instrument Digital Interface. 3, 5, 6, 8–14, 16

**MSD** Million Song Dataset. 8, 11

**WEIRD** Western, Educated, Industrialized, Rich, and Democratic. 15

# 1 Introduction

Music has the ability to convey sentiment and feelings, not only through words and lyrics, but also through the notes and chords themselves. Melodic intervals, chords, tempo, dynamics and timbre all come together to create a certain musical expression that can induce certain feelings in the listener. However, pinpointing exactly what specific part of musical expression induces which kinds of feelings can often be tricky. This is where Artificial Intelligence (AI) and Deep Learning can come into play.

Many applications of AI deal with looking for patterns in data where humans are unable to do so. In the field of Affective Music Composition, AI is used to create new music that is perceived to have a specific emotion or invoke certain emotions in the listener. Possible applications of this include generating soundtracks for movies and video-games in which music plays a big role in setting the emotions of a scene.

While the progress within this field has come a long way the past few decades, extensive work is still to be done, especially within the availability of quality datasets. In this project, a new dataset containing MIDI-files with annotated emotional information is created in order to generate music with a certain emotion using a pre-made generative algorithm. The starting point of this was the interest in improving on the work of Djupvik [Djupvik, 2020] by using a single datasetin the tasks of both music emotion recognition and new music generation, instead of the two vastly different datasets. This project therefore asks the question of whether the creation of a new high quality dataset can improve the performance of Affective Music Composition.

# 2 Background

## 2.1 Generative music and Artificial Intelligence

Algorithmic composition of music has been a growing field ever since it first came into existence in the form of the Illiac Suite, created in 1956 [Hiller, 1981]. The question of whether and how computers can be creative in the context of music has been researched extensively, and has with the rise of AI become a fast-growing subfield in the last decades [Miranda, 2013]. This subfield, often called Musical Computational Creativity (MCC) or Musical Metacreation, is an multidisciplinary research field drawing from AI, art, psychology and cognitive science among others [Sadiku et al., 2019]. The motivation of this research can be largely attributed to the desire to understand the concept of creativity as a whole, the need for generative systems in creative industries such as movies and games, and general computerization of society [Pasquier et al., 2017].

Approaches in MCC can be divided into transformational and generative algorithms, where the former are algorithms transforming an already prepared structure, and in the latter the algorithms create the musical structure themselves [Scirea et al., 2017]. This project makes use of a specific generative algorithm created by Huang and Yang [Huang and Yang, 2020]. This is a generative algorithm that uses the prominent approach of neural sequence models, in which music is considered symbolically equivelent to a language. Musical compositions represented symbolically, meaning as a notation-based format rather audio, are converted into a sequence of words such as Note-On events. In this case, the symbolic music is in the form of MIDI-files. An example of this conversion can be seen in figure 1. A sequence model can then be applied to model the probability distribution of event sequences - in our case, a musical progression - and sample from the distribution to create new music [Huang and Yang, 2020, p. 1].

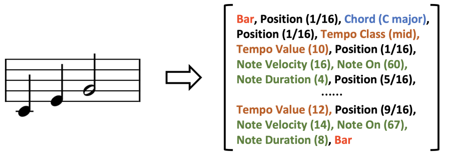

Bar, Position (1/16), Chord (C major),
Position (1/16), Tempo Class (mid),
Tempo Value (10), Position (1/16),
Note Velocity (16), Note On (60),
Note Duration (4), Position (5/16),
......
Tempo Value (12), Position (9/16),
Note Velocity (14), Note On (67),
Note Duration (8), Bar

Figure 1: An example showing the conversion from symbolic music representation into language.

## 2.2 Affective Music Composition

In the search for new ways to enhance the creative performance of MCC, one approach has been to understand and control the emotional expression of the music generated. This is the field of Affective Music Composition (AMC), sometimes also referred to as Affectively-driven algorithmic composition. In most cases, emphasis is put on how real listeners perceive emotions in music, rather than considering induced emotions which could be considered more subjective and influenced by contextual factors [Gómez-Cañón et al., 2021].

This project is related to previous work using music in symbolic form to generate new music with specific emotions. This work often differs in the approach used to represent the emotions expressed by the music. Two predominant taxonomies are heavily used, the discrete or categorical approach, and the dimensional or continuous approach [Gómez-Cañón et al., 2021]. Emotions could be represented within a set of adjectives, such as ”gloomy” or ”serious” [Katayose et al., 1988]. In other work, these adjectives are organized into mood clusters [Hevner, 1936]. These are examples of the categorical or discrete approach.

The dimensional or continuous approach, first proposed by Russel , in which emotions are mapped on to two dimensions [Russell, 1980]. The terms used for these dimensions differ, but is most commonly referred to as **valence and arousal** [Scirea et al., 2017]. Valence relates to the pleasantness or positiveness of an emotion, while arousal relates to the energy or activation. This two dimensional space can be visualised, and areas labeled with categorical emotions. An example of this can be seen in figure 2, taken from the work of Ferreira and Whitehead [Ferreira and Whitehead, 2019]. In this project, the perceived emotions of music were labeled using the continuous approach.

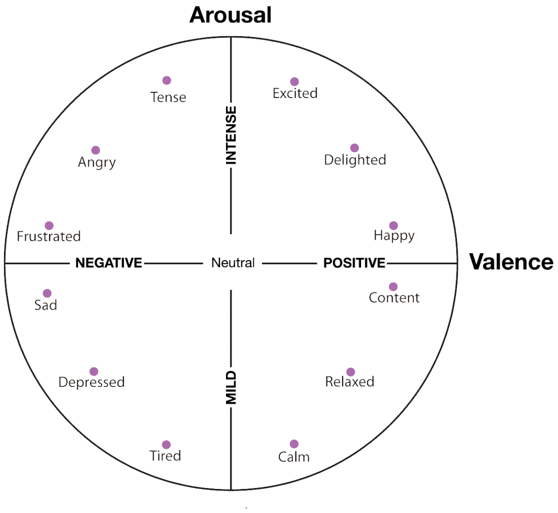


Figure 2: The valence-arousal space with categorical labels.

## 2.3 Existing datasets

One of the prominent challenges in the field of Musical Computational Creativity in AI is the availability of high quality datasets. Especially within AMC and MER, most datasets are either small, insufficiently diverse in terms of genre, or not open to the public. Panda, as well as Gómez-Cañón et al., provides good overviews of existing datasets, highlighting the shortages of these [Panda, 2019, Gómez-Cañón et al., 2021]. While significant progress has been made in the field in later years, and new datasets have been created, such as VGMIDI [Ferreira and Whitehead, 2019], efforts to find a dataset of significant size and genre diversity for this project were unsuccessful. Metacreation provides a an extensive list of available collections of audio and MIDI files for use in AI research[1].

In this project, some existing musical datasets were used in testing and creation of the new dataset. Figure 1 shows an overview and short description of these.

| Dataset name | Creator | Year | Contents |
|---|---|---|---|
| ADL Piano MIDI | Ferreira et al. | 2020 | 11 086 MIDI files of piano pieces from different genres |
| Million Song Dataset | Bertin-Mahieux et al. | 2011 | Audio features and metadata for a million contemporary popular music tracks |
| Deezer Mood Detection Dataset | Delbouys et al. | 2018 | Valence and arousal annotations for 18 645 of the songs from the Million Song Dataset |

Table 1: The existing datasets used in this project

## 2.4 Dataset creation

Gómez-Cañón et al. outlines a methodology of creating new datasets for use in MER and AMC, as well as highlighting the obstacles and trade-offs in doing so [Gómez-Cañón et al., 2021]. Collection of emotion annotations is usually performed by collecting emotional evaluations from one or more listeners, preferably of a variety of cultural and demographic backgrounds. The average or median of these annotations are then used as a ground truth, although this is a simplification of a very complex concept. A more complex approach also takes the variation in cultural and demographic background into account in the data collection, creating personalized models. Gómez-Cañón et al. (ibid) further recommend collecting both categorical and continuous emotional information, as well as collecting user data such as musical background.

## 2.5 Outset and refining of project

The starting point of this project was the hope of refining some of the work of Djupvik [Djupvik, 2020], in which the end goal is much the same as in this

[1] https://metacreation.net/corpus-1/

project. Djupvik approaches the issue in two different steps First, Music Emotion Recognition (MER), another field in AI, is used in order to recognize emotions in MIDI-files in an automated manner. These annotated MIDI-files are in turn used to generate music with the same recognized emotion. One issue with this approach is that one can not completely trust the annotations that the MER algorithm makes, especially as the dataset used for training the MER algorithm is very different to the music that it is then asked to annotate. Djupvik uses a dataset of popular music of multiple genres in her MER-algorithm, and the MASESTRO dataset, which only consists of classical piano music, for the composition training [Hawthorne et al., 2019]. This results in that the MIDI-files in the composition dataset are all recognized as having low intensity [Djupvik, 2020, p. 66].

In trying to improve on this, the idea is to use the same dataset for both tasks. However, such a dataset would have to consist of songs in MIDI-format with annotations for arousal and valence, which is precisely what is obtained in the MER part of the process. Thus, having this dataset would eliminate the need to use this algorithm in order to achieve the goal of generating affective music. One could however also use this same dataset for MER-tasks, and could in this project have been used in order to expand the dataset used for composition by obtaining emotion annotations on MIDI-files not yet annotated. However, this was deemed not necessary and out of the scope of this project.

# 3 Method

The project is conducted in three main phases. The first phase is obtaining a collection of MIDI-data that can be used for training the algorithm. The second is collecting valence and arousal information on each of the MIDI files. Finally, the files and data is used in training the algorithm output music that expresses specific emotions.

A minimum required number of MIDI-files is a number between 500 and 2000. These files is obtained through searching for existing collections of files, and selecting a subset of these files. This selection is based on a series of criterion, described as follows:

- Due to limitations of the AI algorithm, music with note values lower than sixteenth-notes are unfavourable, as the algorithm will quantize any notes faster than this into sixteenth-notes.
- It is favorable that the music features some form of melody, as apposed to only chords, so that the algorithm will have more room for expression.
- Files containing large segments of silence are filtered out.
- The music has to be notated more or less "correctly", meaning that it is written in the correct time signature, that the music matches the indicated tempo, and otherwise is not written in such a way that the notation would be deemed incorrect.

In the second phase, valence and arousal values for these files are obtained. This is done using human participants listening to the music and rating the emotional content of each song. The MIDI-files are therefore rendered into digital audio files using some computer software. As there are many files to annotate, each participant recieves a smaller, random selection of the files, but are encouraged to answer as many as they would like. A web based survey solution is used in order to reach the participants without the need for physical meetups.

Finally, the files and data are used in training the algorithm, which in turn produces music that expresses certain emotions. Using the valence and arousal data, we can divide the dataset into four quadrants corresponding to high and low values of valence and arousal, as seen in figure 2. The musical generative algorithm is then trained on the collection of songs within each quadrant separately. The result will be four models, which then can be used to generate music expressing the emotions corresponding to the respective quadrant.

The generative algorithm used is the one created by Huang and Yang, using the instructions provided with their code[2] [Huang and Yang, 2020]. All other code related to the creation of the dataset will be written in Python.

[2] https://github.com/YatingMusic/remi

# 4 Results

## 4.1 Obtaining MIDI-files

One of the most difficult challenges was to obtain a usable set of MIDI-files. As discussed in section 2.3, several datasets of MIDI-files already exist, but they all have their limitations. Such limitations are either in terms of size, quality, availability or the diversity of the music. For example, the MASESTRO dataset is commonly used in the music generation purposes, but only contains classical music, and does not contain any information on the musical tempo [Hawthorne et al., 2019]. Other collections of MIDI-data are of extremely varying quality, but in the end, the ADL Piano MIDI dataset was chosen as the basis of the new dataset. This dataset prevails mainly because of its diversity of genres and size, as well as its link to the Million Song Dataset (MSD), which contains useful metadata [Bertin-Mahieux et al., 2011]. The MIDI-files were matched with songs in MSD and renamed to a corresponding MSD unique track ID. Because of limitations in the generative software, all files were converted to a format where all MIDI-events were in the same MIDI-channel, track and instrument. The files were then manually listened to, and 500 files were selected for use in the dataset.

## 4.2 Testing

Before obtaining sentiment information, some testing was performed in order to ensure that the files were usable and of sufficient quality for this task. Also, the testing proved valuable in learning about how the training data influences the generated music.

Initially, the intention to use MIDI-files with more than one instrument. This was desirable because it would give the music more tools for expression and would therefore be able to convey a more discernible emotion. However, the software used is written to only work with one instrument, and efforts to adapt the software were deemed outside the scope of this project due to time constraints. A search for alternative software that could do this was done, but unfortunately no free, openly available option was found. Thus, the project was constrained to only working with piano music.

A test dataset was made in order to test the software and learn more about how it works and how to maximize its performance. This dataset was created by matching the music to MSD, and then matching these entries to the Deezer Mood Detection Dataset [Delbouys et al., 2018]. This is a dataset containing valence and arousal values for a selection of entries in MSD. However, a notable issue with this dataset is that the emotion values are derived from associated written tags taken from LastFM[3]. These annotations are therefore not very trustworthy, both because they are not directly mapped by human participants, and also that they are derived in relation to the full audio version of the song rather than its MIDI representation. These versions can obviously be very

[3] https://www.last.fm

different from each other, and could therefore convey entirely different emotions. However, testing with this dataset was deemed a useful step in the process as it could give some insight of which kind of results that could be expected. It also helped establishing the criterion for the selection of MIDI-files discussed in section 3.

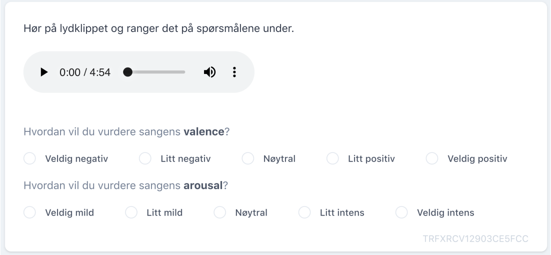
The results from generation using the test dataset were discouraging. The music could not be said to convey the specified emotions, and was in general incoherent and lacking musical structure.

Another test, in which the software was given an extremely homogeneous set of MIDI-files, was also conducted. Eighteen versions of the same song, the norwegian children's tune ”Lisa gikk til skolen”, was created, only differing in tempo and key, and used to train the AI. Results from this were far more encouraging, as the algorithm composed songs of similar structure and mood as the original music, while also being creative in changing up the melody and chord progression.

## 4.3 Obtaining sentiment information

Values for arousal and valence of the MIDI-files was obtained using voluntary human participants. A website app[4] was created in which the subject was presented with a brief description of the project, explanation of the concepts of valence and arousal, and twenty audio files randomly chosen from the 500 files in the dataset. Each participant was also encouraged to complete the survey several times, with a new 20 songs each time.

Participants were instructed to rate each song on a five point scale for valence and arousal, in which valence went from very negative to very positive, and arousal went from very mild to very intense. Figure 3 shows an example of a question in the survey.



Hør på lydklippet og ranger det på spørsmålene under.

Hvordan vil du vurdere sangens **valence**?

○ Veldig negativ ○ Litt negativ ○ Nøytral ○ Litt positiv ○ Veldig positiv

Hvordan vil du vurdere sangens **arousal**?

○ Veldig mild ○ Litt mild ○ Nøytral ○ Litt intens ○ Veldig intens

TRFXRCV12903CE5FCC

Figure 3: An example of a question from the survey.

The survey was distributed to friends, acquaintances and family, which were also encouraged to share it forward to their own social groups. In total, the survey collected 1180 responses. Due to the files being randomly selected for each participant, the number of evaluation per file also varies, with some not

[4] https://music-survey.vercel.app/

receiving any response at all. These were manually sent to a few participants for rating, in order to ensure all 500 files received at least one evaluation. Figure 4 the distribution of responses over the files. Of the 500 files, 174 of them only received one In cases where a file had more than one response, the values were averaged.

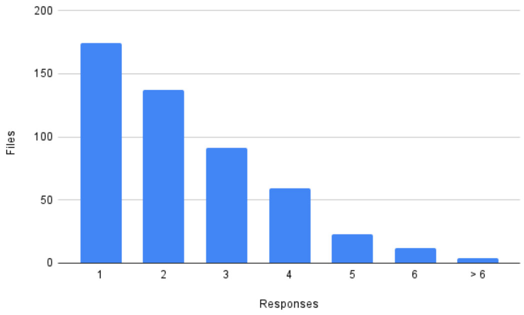


Figure 4: The distribution of number of evaluations collected for each file.

With sentiment information gathered, the dataset was complete, consisting of 500 MIDI-files, and a CSV file containing the valence and arousal values of these files. The values were normalized to be between -1 and 1, and the dataset was divided into the four quadrants of the valence-arousal plane by dividing on whether the values were above below ”neutral”. Files that had a value of 0 in either valence or arousal were not included in any quadrant. Table 2 shows the quadrants and their corresponding values of valence and arousal, as well as the amount of files sorted into each quadrant. We can see that there is an uneven number of files in each quadrant, with Q4 having around twice as many files as the other quadrants, and Q2 having the fewest.

| Quadrant | Valence | Arousal | No. of files |
|---|---|---|---|
| Q1 | Low | Low | 72 |
| Q2 | High | Low | 55 |
| Q3 | Low | High | 76 |
| Q4 | High | High | 134 |

Table 2: The quadrants of the valence-arousal plane.

## 4.4 Training and music generation

Further, the MIDI-files from each quadrant were used to train one instance of the AI algorithm each, in order to obtain four different models corresponding to each quadrant. Training was allowed to continue until complete, taking 28 epochs for Q1, 30 epochs for Q2, 26 epochs for Q3 and 41 epochs for Q4.

Finally, the models obtained from training were used to generate new music. Nine different songs of 16 bars were generated from each of the models, totalling in 36 songs. MIDI and MP3 versions of these songs are provided as attachments.

## 4.5 Description of generated songs

The quality of the music produced is in general highly varied and inconsistent. However, some characteristics stood out more often than others. Much of the generated music contained rapid, staccato sixteenth notes, regardless of which quadrant the training data was taken from. In some cases, the algorithm would seem to get "stuck" on a chord, never switching away from it throughout the song. An example of these features can be seen in the sheet music in figure 5.

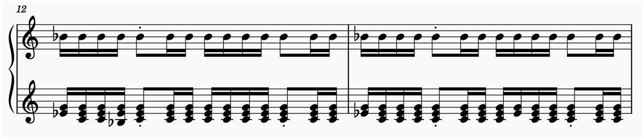


Figure 5: An excerpt from the sheet music for composition "Q4-temp=0.8-topk=4"

In other cases the algorithm would compose coherent, seemingly well thought out musical pieces. Figure 6 shows an example of a composition containing a coherent song structure, an interesting chord progression, as well as recurring melodic themes throughout the music (although sometimes poorly notated). Compositions such as this one were few and far between, but at least shows that the algorithm has the ability to create coherent music.

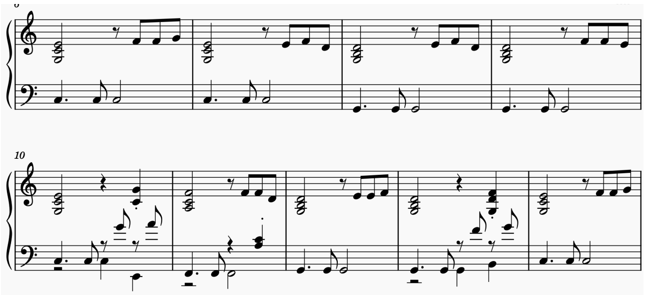


Figure 6: An excerpt from the sheet music for composition "Q1-temp=1.2-topk=4"

# 5 Discussion

## 5.1 Resulting dataset

The main focus of this project was the creation of a substantial dataset of MIDI-files with corresponding sentiment annotations, as few such datasets are available. In that regard, the dataset produced here is arguably a useful contribution to the field of Affective Music Composition. The dataset is larger than existing datasets of the same nature, and contain a larger diversity in terms of music genres.

However, the dataset also has its limitations and flaws. The main limitation is still with regards to size. While larger than other similar datasets, the created dataset is still not as large as is often necessary in MER and MCC applications. This limitation was simply set by the need to manually listen to each of the files in order to evaluate the quality and the suitability of the data. This task is very time consuming and tedious, which has directly limited the amount of files included in the dataset. Future work could expand on this by simply working with more files.

Another flaw of this dataset is the way in which sentiment data was obtained. In order to make the task of annotating the data as simple as possible, and in the hopes of maximizing the amount of responses, only valence and arousal data was collected from the participants. Gathering of categorical data, such as having the participants write or choose from adjectives describing the music, would have made the dataset more useful for different use cases. In addition, valence and arousal annotation was done on a five point scale, limiting the options and eliminating some nuance between perceived emotions. A different approach could incorporate sliders, or even have the participants choose a point on the valence-arousal plane, allowing for more more precise annotation. However, this could also lead to more confusion and tediousness of the task, possibly reducing the number of responses.

Even with the efforts made to maximize the amount of responses, the number of evaluations for single tracks is still a weakness of the resulting dataset. With a large portion of the music only receiving a single evaluation, the data is not very reliable, as it only represents one participant's interpretation. A larger number of evaluations per file improves the objectivity of the sentiment data. In order to achieve this, rather than using randomization, one could have given the files to the participants in a certain order while still ensuring a participant does not receive the same file more than once. This was not done due to technical and time limitations. It would also be preferable to have more participants in order to improve objectivity.

The quality of the responses can also be called into question. The participants were taken from a small subset of the population, consisting only of people in or just outside of the author's social circle. This led to the vast majority of the participants being Western, Educated, Industrialized, Rich, and Democratic (WEIRD), which can have implications on the generalization of the results to a wider audience [Gómez-Cañón et al., 2021]. Furthermore, the degree of understanding of the task each participant had can not be established. The concepts of valence and arousal are not straightforward and may be subject to personal interpretation, and this could vary greatly from person to person. One way to counteract this could be to have a more in-depth explanation, or have musical examples explaining the terms, though this also has its negative implications on number and quality of the responses. Overall, many trade-offs were made in the effort of receiving as many responses as possible in the short allotted time span. Better data reliability and quality could be ensured through efforts in terms of demography, explaining terms, and expanding on how the respondents can express emotions they experience in the music.

## 5.2 Generated music

As discussed in section 4.5, the quality of the compositions generated by the algorithm was very inconsistent. Its tendency to use rapidly repeating sixteenth-note chords makes the music at times seem very aggressive and unpleasant. These types of compositions are certainly of high arousal, but will sometimes be hard to discern whether they are positive or negative - they are rather just messy. No conclusive reason for this phenomenon could be found in this project, as the training data does not prominently contain music of this nature. The tendency to choose high tempos, however, could be attributed to the fact that tempo in MIDI-files is usually only mentioned once in the beginning of the song, which could downplay the importance of this variable to the algorithm.

However, in the instances where the algorithm is able to create more controlled, coherent music, the results are more impressive. Following is a description of some musical features found in a selection of the produced material for each quadrant.

### Q1 - Low valence, low arousal

For Q1, instances such as ”Q1-temp=1.2-topk=4” (see figure 6) features what can be described as a simple and mild ballad, although the valence of the composition could be argued to be positive. The chord progression of ”Q1-temp=0.8-topk=4” is perceived as negative, and could be be said to have low arousal as well. ”Q1-temp=1.0-topk=6” starts off sounding very calm and tired, before incorporating some unusual chords towards the end.

### Q2 - High valence, low arousal

The generated music trained on MIDI from Q2 can be said to contain generally positive emotions. However, the model struggles when it comes to containing the arousal of the music. Even when ignoring the pieces with rapid sixteenth-notes, most of the music is fairly energetic, suggesting it should rather belong to Q4. One of the produced pieces, ”Q2-temp=1.0-topk=6”, does a better job of expressing the correct emotion, although the tempo is a little high to be described as low arousal.

### Q3 - Low valence, high arousal

The model for Q3 seems to produce music which often does not fit the corresponding emotions. ”Q3-temp=1.2-topk=5” features a familiar chord progression in a minor key, with high tempo and melodic movement, sounding both negative and energetic. ”Q3-temP=1.2-topk=6” and ”Q3-temp=1.0-topk=5 has high arousal, but valence seems positive. ”Q3-temp=1.0-topk=4” starts off sounding anxious, perhaps being best placed in Q1, but evolves into a fast chord progression that sounds very happy and energetic. One reason for this quadrant performing poorly could be that low valence, high arousal music is not usually found in piano music.

### Q4 - High valence, high arousal

The music generated from the Q4-model contains ”Q4-temp=1.0-topk=4” featuring an upbeat rhythm with a mostly positive chord progression, and even including a bright melody at one point. ”Q4-temp=0.8-topk=5” is also a very happy melody, although perhaps in a little high tempo. ”Q4-temp=0.8-topk=6” has a high energy bass-line, but becomes quite repetative and lacks structure.

### Overview

Overall, we see slight success in achieving the goal of generating affective music. The resulting compositions is an improvement compared to the results obtained by Djupvik using the same generative algorithm [Djupvik, 2020]. This suggests the proposed method of improving the performance of the algorithm does work to a degree, but is still not as successful as originally hoped. A lot of variables are involved, and it is difficult to pinpoint exactly what causes the behaviour. It could also be argued that some degree of picking and choosing among the produced material should be accepted, as it also is for human composers not releasing anything they produce. However, an even higher quality dataset, more insight into how the algorithm works in order to better finetune the training data, or using a different algorithm altogether could be some of the areas of improvement.

# 6 Conclusion

In conclusion, the work done in this project have been useful and productive to the field of Affective Music Composition in that a new dataset has been made available for future use. The dataset consists of 500 openly available MIDI songs, annotated with valence and arousal data and corresponding quadrant belonging. Also, the songs included correspond with the MSD dataset, allowing for further combination with other metadata if desired in other projects. However, the quality of this dataset is certainly questionable in some regards, and could definitely have drawn benefits from expansion in terms of size and annotations, as well as the amount of responses for each MIDI-file. The majority of the work needed to improve these shortcomings is tedious, and not easily automated, but can with more time and effort easily be done.

The music generated by the algorithm shows promise, and is a proof of concept for this approach to generating affective music. It does, however, need some work in improving the consistency of the algorithm, and eliminating the production of music with extremely rapid notes.

This project has been interesting as well as challenging. Often the challenges arose in parts of the project thought to be straightforward, but in the end, the approach in which the result was achieved was one I was happy with. The knowledge gained from researching this topic has been very instructive and interesting, and has given a lot of inspiration for future work within the field of AI.

# References

[Bertin-Mahieux et al., 2011] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011).*

[Delbouys et al., 2018] Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., and Moussallam, M. (2018). Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276.*

[Djupvik, 2020] Djupvik, A. (2020). Music that feels just right: Emotion-based music classification and composition. Master's thesis, NTNU.

[Ferreira and Whitehead, 2019] Ferreira, L. N. and Whitehead, J. (2019). Learning to generate music with sentiment. *CoRR*, abs/2103.06125.

[Gómez-Cañón et al., 2021] Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., and Gómez, E. (2021). Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Processing Magazine*, 38(6):106–114.

[Hawthorne et al., 2019] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J., and Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations.*

[Hevner, 1936] Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268.

[Hiller, 1981] Hiller, L. (1981). Composing with computers: A progress report. *Computer Music Journal*, 5(4):7–21.

[Huang and Yang, 2020] Huang, Y.-S. and Yang, Y.-H. (2020). Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

[Katayose et al., 1988] Katayose, H., Imai, M., and Inokuchi, S. (1988). Sentiment extraction in music. In *9th International Conference on Pattern Recognition*, pages 1083–1084. IEEE Computer Society.

[Miranda, 2013] Miranda, E. R. (2013). *Readings in music and artificial intelligence*. Routledge.

[Panda, 2019] Panda, R. E. S. (2019). *Emotion-based analysis and classification of audio music.* PhD thesis, 00500:: Universidade de Coimbra.

[Pasquier et al., 2017] Pasquier, P., Eigenfeldt, A., Bown, O., and Dubnov, S. (2017). An introduction to musical metacreation. *Comput. Entertain.*, 14(2).

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

[Sadiku et al., 2019] Sadiku, M. N., Ampah, N. K., and Musa, S. M. (2019). Computational creativity. *Lecture Notes in Computer Science*, 3(6):1–3.

[Scirea et al., 2017] Scirea, M., Togelius, J., Eklund, P., and Risi, S. (2017). Affective evolutionary music composition with metacompose. *Genetic Programming and Evolvable Machines*, 18(4):433–465.

# Attatchments

1. *code* - The self-produced code used in the project, as well as the code taken from Huang and Yang, contained in the folder *remi*.
2. *dataset* - The dataset containing MIDI-files and a .csv file with valence and arousal values
3. *models* - The models from training on the dataset
4. *music* - MIDI-files containing the music produced by the algorithm

NTNU

Kunnskap for en bedre verden