CS4347 Project Proposal: Towards Empathetic Music Generation System

Poh Lin Wei, Ye Chenchen, Hu Yidi
A0190339B, A0177915R, A0173378U

1 Motivation

In the last few years, research in music generation systems has gained traction. For instance, in 2016, Google announced Magenta which is a research project that explores the use of machine learning in creating music and art [1, 2]; in 2017, the singer Taryn Southern released an album which was produced with the help of automated music generation systems [14]; and in 2019, OpenAI announced MuseNet which can compose 4-minute musical pieces in different styles [3]. Despite these advancements, relatively lesser attention has been paid to developing systems that allow listeners to control the emotions expressed in the generated music pieces.

This is regrettable as such systems have useful applications, like in music therapy and/or music medicine for individuals with depression. Specifically, under certain circumstances, the use of new and unfamiliar generated music during therapy sessions may be preferred since it can provide the client with a new perspective [15]; and avoid triggering unpleasant memories [23, 10]. With that said, merely being able to generate music may not suffice in some situations. Instead, it will be more desirable if the user could generate music which expresses a specific emotion. The generated music may then be used to alter the listener's mood [17]; or as a medium through which he or she can explore and experience the specified emotion [9]. Given these considerations, attempts to develop systems that can generate music, with emotional content which matches that selected by the listener, are undoubtedly constructive.

Therefore, we propose looking into possible methods to creating such systems.

2 Goal and Objectives

We propose exploring alternative approaches to creating an empathetic music generator: given an emotion label, it should be able to generate music which expresses the specified emotion.

Considering that multiple studies have explored the use of handcrafted rules to develop such a system [28, 21, 20], and that it is difficult, and even impossible, to enumerate all the rules such that the result is not only satisfactory but also diverse, we would like to focus on creating such a music generator with minimal hand-crafted rules.

In addition, we propose to solely focus on the generation of single-instrument music pieces, since the focus is on exploring possible approaches to creating a music generator that is empathetic rather than on exploring possible methods to generating music in general. Specifically, we believe that generation of keyboard pieces is a good option due to the availability of datasets for the keyboard, and the fact that it is a flexible instrument that is capable of playing multiple notes at once and a wide-range of notes.

Furthermore, since the purpose of the empathetic music generator is to generate music which expresses a specified emotion, the ultimate output of the generator should not merely be music notations because the way in which these notations are performed is equally important for a more in-depth discussion). With that said, it might not be desirable to create a music generator that produces raw audio because modelling raw audio is not only very challenging but also a computationally expensive task [8]. Thus, a good compromise will be the approach taken by Oore et al. [25]: the system should output a MIDI representation of a musician's performance (as opposed to a musical score).

Lastly, since the study is largely an exploratory one, we believe that it will suffice to generate music pieces which are capable of expressing high/low valence and high/low arousal, or vice versa. In other words, we will discretise the valence-arousal (VA) representation of emotion into four

categories: high valence and low arousal, low valence and low arousal, high valence and high arousal, and low valence and high arousal. This choice of emotion representation also allows us to use a wider range of datasets to train the system.

3 Solution

3.1 Overview

Given the limitations of incorporating and relying on handcrafted rules to achieve the desired result, we hope to explore the development of systems that will not only learn to generate music but also learn the relationship between music and emotions. Furthermore, in view of the success of deep learning techniques in various generative tasks [5, 27, 3], we would like to adopt such an approach. More specifically, to overcome the main limitation of Monteith et al. [24]'s work, in which multiple pieces of music might have to be generated before a suitable piece is selected as the system's final output, we propose examining the feasibility of using class-conditioned deep learning models to create an empathetic music generator.

3.2 Music Representation

The music representation that will be used will largely be similar to Oore et al.'s [25].

In their work, each time step is defined using absolute time interval. As compared to defining it based on the number of musical beats, their approach is advantageous since it can easily capture a musician's use of musical devices like rubato (in which a musician would intentionally slow down or speed up for dramatic effect, and hence the notes may not nicely align with the specified musical beats).

Furthermore, since the dataset that will be used to train the model is in the form of MIDI files and the music generator's final outputs should be MIDI files, the representation used should be compatible with MIDI.

A MIDI file is made up of a sequence of MIDI events, where each event indicates which and how the 128 musical notes are affected (e.g. the notes' loudness and how long they are played for). Therefore, Oore et al. [25] suggested using the following representation when training the model. Each input and output should be a one-hot 413-dimensional vector: for each of the 128 pitches, 1 dimension is used to represent a note-on event (which indicates that the note is a new note to be played) and 1 dimension is used to represent a note-off event (which indicates that the note should be released); 125 dimensions indicate the duration (in milliseconds) of the current event; and the last 32 dimensions indicate the loudness with which the notes should be played.

3.3 Music Generation System

In light of the promising potential of music generation using neural networks [25, 18, 26], we intend to adopt this approach as well.

Ideally, we would use a state-of-the-art architecture to generate music. However, given the computational costs required by such models and the exploratory nature of this proposed study, we believe that a compromise should be made between the cost of training the model and the quality of the generated music. Since the state-of-the-art architectures are based on Transformer models [31], which use an encoder-decoder structure and attention, and several works like Oore et al.'s [25] have shown that LSTM-based RNNs are capable of generating satisfactory abstracts of music, sequence-to-sequence models with attention [6] would be a good compromise. Hence, we intend to adapt Bahdanau et al.'s [6] architecture for the purpose of music generation.

To condition our music generator, the first and simplest approach is similar to that adopted by Ghosh et al. [16] and Alvarez-Melis et al. [5]. Instead of simply providing the generator with the output from the previous time step to generate the next output, we would also provide a vector, or an emotion label, that represents the desired emotion which should be expressed by the model. In this approach, at each time step, we would concatenate the emotion label with the previous output and pass this concatenated vector into the model.

The second approach to conditioning is inspired by the works like Riecken [28], Herberger et al. [30] and Lucas et al. [21], which defines a fundamental music unit and associating it which a specific emotion, and also works like Monteith et al. [24] and Legaspi et al. [20], which allows

system to learn musical features and elements that can express a given emotion. A cursory glance at those approaches would reveal that they are built on the premise that certain musical features would express certain emotions. Therefore, if we could select the "right" musical features and use them for generating music, the resulting piece should express the desired emotion. Unfortunately, as suggested previously, it is challenging to manually specifying rules to select combinations of musical features that are not only suitable but also varied.

Given these observations, we feel that an intuitive solution is to teach the system to generate musical features which are congruent with the user-selected emotion. Then, these features can be used to generate music. Intuitively, this approach seems more aligned with the approach taken by composers because ultimately, the musical features are the tools that are used to create music, not the emotions. How could we then generate these features? We propose doing so using a conditional variational autoencoders (CVAE). Specifically, before training the CVAE, we would extract musical features from a dataset which has music pieces that are labelled with emotions. The choice of musical features to be extracted is informed by studies which examine the relationship between music and emotions [13, 29, 32]. Using the extracted musical features and the associated emotion label, we can train the CVAE.

In summary, in the second approach, the system uses two generators: one feature generator, which would create the desired musical features, and one music generator, which would be conditioned on these features to create a piece of music.

Regardless of the approach taken to condition the music generator, the music generator would be trained using the curriculum learning approach proposed by Huang et al. [19]. As opposed to teacher forcing, which was used in previous works [25, 18], their approach is able to ameliorate the issue of exposure bias. When using teacher forcing to train a model, the model will not be exposed to its own errors because at each time step, the true previous-step output is fed into the model to generate, not its predicted previous-step output. However, this is not the case during training. This difference in training and testing conditions is known as exposure bias [7].

4 Resources and Timeline

4.1 Dataset

The dataset which will be used to train the music generator is the Piano e-Competition dataset [4] which consists of at least 1400 performances in MIDI representation. As these performances were conducted by skilled pianists, the model could potentially learn to generate expressive music with appropriate dynamics and musical expressions. To augment the data, we intend to use the Oore et al.'s approach [25]. Specifically, given a piece of music, we will alter the pitch of every note by the same amount, or slow it down or speed it up. The former form of data augmentation should not significantly affect the quality of the music since we often perceive pitch relatively rather than absolutely [22]. Similarly, the latter would not degrade the data's quality since the music would either sound slightly slower or faster.

Unfortunately, as the music pieces from the Piano e-Competition dataset are not labelled with emotions, we would need to use the VGMIDI dataset [12] to train a model to label them. (We do not think that it is wise to train the music generator using the latter dataset as it is significantly smaller than the former. Hence, the additional step of training a classifier that can label the former dataset.)

In the event that the data we have is insufficient, we would use the ADL Piano MIDI Dataset [11] to train an initial model that can be fine-tuned using the aforementioned datasets.

4.2 Timeline

To ensure that we are able to get the project completed within the given timeframe, we proposed a project timeline with milestones set in weeks. Firstly, starting from week 1, we will mainly focus on data preparation for the model training. Specifically, we plan to finish training the labeling system on the VGMIDI dataset, and apply this system to label music pieces from the Piano e-Competition dataset and the ADL Piano MIDI dataset with emotions. The data preparation task shall be completed in 1.5 weeks. Secondly, from week 3 to week 5, our plan is to construct and train the two generator model separately: the feature generator that encodes emotion label and music features to music-emotion features, and the music generator that takes in the music-emotion

features and music primers to form generate music that expresses the given emotion. Finally, we will integrate these two models and fine-tune the parameters using the dataset we labeled in week 1. A final report will also be developed and delivered by week 6, where more details about our motivation, research review, experiment, and analysis will be included.

References

[1] Google has set up an ai group called 'magenta' to see if computers can produce original art and music.

[2] Make music and art using machine learning.

[3] Musenet.

[4] Piano e-competition.

[5] David Alvarez-Melis. The emotional gan : Priming adversarial generation of art with emotion. 2017.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[7] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, 2015.

[8] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer, 2018.

[9] Tan-Chyuan Chin. *Measuring adolescents' emotional responses to music: Approaches, challenges, and opportunities*, page 39–52. Oxford University Press, May 2019.

[10] Kerstin Denecke. A mobile system for music anamnesis and receptive music therapy in the personal home. *Studies in health technology and informatics*, 245:54–58, 2017.

[11] Lucas N Ferreira, Levi HS Lelis, and Jim Whitehead. Computer-generated music for tabletop role-playing games. 2020.

[12] Lucas N. Ferreira and Jim Whitehead. Learning to generate music with sentiment. 2019.

[13] Aalf Gabrielsson and Erik Lindström. *The Role of Structure in the Musical Expression of Emotions*, page 367–400. Oxford University Press, Jul 1993.

[14] Dom Galeon. The world's first album composed and produced by an ai has been unveiled.

[15] Susan C. Gardstrom and James Hiller. Song Discussion as Music Psychotherapy. *Music Therapy Perspectives*, 28(2):147–156, 11 2010.

[16] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[17] Annie Heiderscheit and Amy Madson. Use of the Iso Principle as a Central Method in Mood Management: A Music Psychotherapy Clinical Case Study. *Music Therapy Perspectives*, 33(1):45–52, 02 2015.

[18] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer, 2018.

[19] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, and Mi Zhang. Dance revolution: Long sequence dance generation with music via curriculum learning, 2020.

[20] Roberto Legaspi, Yuya Hashimoto, Koichi Moriyama, Satoshi Kurihara, and Masayuki Numao. Music compositional intelligence with an affective flavor. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, page 216–224, New York, NY, USA, 2007. Association for Computing Machinery.

[21] Pedro Lucas, Efraín Astudillo, and Enrique Peláez. *Human–Machine Musical Composition in Real-Time Based on Emotions Through a Fuzzy Logic Approach*, page 143–159. Springer International Publishing, Oct 2016.

[22] Josh H McDermott and Andrew J Oxenham. Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology*, 18(4):452–463, Aug 2008.

[23] B. Medcalf. Exploring the music therapist's use of mindfulness informed techniques in practice. *Australian Journal of Music Therapy*, pages 47–66, 2017.

[24] Kristine Monteith, Tony R. Martinez, and Dan Ventura. Automatic generation of music for inducing emotive response. In *ICCC*, 2010.

[25] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: learning expressive musical performance. *Neural Computing and Applications*, 32(4):955–967, Nov 2018.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[27] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019.

[28] D. Riecken. Wolfgang: "emotions" plus goals enable learning. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, volume 2, pages 1119–1120 vol.2, 1998.

[29] Yading Song, Simon Dixon, and Marcus T. Pearce. Evaluation of musical features for emotion classification. In *ISMIR*, 2012.

[30] Titus Tost Tilman Herberger. System and method of automatically creating an emotional controlled soundtrack, U.S. Patent 7,754,959B2, Dec. 2005.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[32] Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4):365–389, Aug 2017.