UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CONTROLLING PERCEIVED EMOTIONS IN COMPUTER-GENERATED MUSIC

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Lucas N. Ferreira

July 2021

The Dissertation of Lucas N. Ferreira
is approved:

Professor Jim Whitehead, Chair

Professor Marilyn Walker

Professor Levi Lelis

Dean John Doe
Vice Provost and Dean of Graduate Studies

Chapter 1

Introduction

Music composers have been using algorithms, rules and general frameworks for centuries as part of their creative process to compose music [30]. For example, Guido of Arezzo (around 991-1033), in his work "Micrologus", describes a system for the automatic conversion of text into melodic phrases. French composers of the *ars nova*, such as Phillipe de Vitry (1291–1361) and Guillaume de Machaut (1300–1377), used isorhythms as a method to map a rhythmic sequence (called *talea*) onto a pitch sequence (called *color*). In the "The Art of the Fugue", Johann Sebastian Bach (1685–1750) deeply explores contrapuntal compositional techniques such as the fugue and the canon, both being highly procedural.

Since the 1950s, scientists, engineers and musicians have been designing algorithms to create computer programs capable of composing music automatically. The "ILLIAC Suite" is considered to be the first piece to be fully composed automatically by an electronic computer [30]. The program that generated this composition was written by Lejaren Hiller and Leonard Isaacson for the ILLIAC computer at the University of Illinois [18]. Since then, many different

methods have been proposed to generate music with computers: rule-based systems [], grammars [44, 38], evolutionary algorithms [20], learning-based systems [17], etc. The scientific (and artistic) field that organizes these algorithms is called Algorithmic Music Composition. The work in this field has influenced different music genres such as Generative Music [10] and supported applications in music analysis [27], procedural audio [11], audio synthesis [9], music therapy [48], etc.

With the great advances in Neural Networks since the 2000s, Deep Learning methods achieved impressive results in many areas of Artificial Intelligence (AI) such as Computer Vision (CV), Speech Recognition and Natural Language Processing (NLP) [13]. Consequently, Algorithmic Music Composition researchers started exploring different types of neural networks to generate music: Recurrent Neural Networks (RNNs) [31], Transformers [21], Convolutional Neural Networks (CNNs) [50], Variational Autoencoders (VAEs) [40], Generative Adversarial Networks [8], etc. One of the most common approaches of neural-based music generation consists of using a Transformer (or a RNN) to build a neural music language model (LM) that computes the likelihood of the next musical symbols (e.g., note) in a piece. Typically, the symbols are extracted from MIDI or piano roll representations of music [3]. Music is then generated by sampling from the distribution learned by the LM. Such models have been capable of generating high quality pieces of different styles with strong short-term dependencies. Supporting strong long-term dependencies (e.g. music form) is still an open problem [3].

A major challenge of music LMs consists of disentangling the trained models to generate compositions with given characteristics [12]. For example, one cannot directly control a model trained on classical piano pieces to compose a tense piece for a horror scene of a movie.

Being able to control the output of the models is especially important for the field of Affective Algorithmic Music Composition, whose major goal is to automatically generate music that is perceived to have a specific emotion or to evoke emotions in listeners [47]. Applications involve generating soundtracks for movies and video games [46], sonification of biophysical data [4] and generating responsive music for the purposes of music therapy and palliative care [28].

One of the most common approaches to Affective Algorithmic Music Composition is rule-based systems [47], which use rules encoded by music experts to model principles from music theory to control the emotion of generated music. These systems are helpful in systematically investigating how a small combination of music features (tempo, melody, harmony, rhythm, timbre, dynamics, etc.) evoke emotions. However, due to the large space of features, it is challenging to create a fixed set of rules that consider all music features. Learning-based methods do not have this problem because one does not need to specify the rules to map music features into emotion. These rules are learned directly from music data.

This dissertation explores how to control LMs to generate music with a target perceived emotion. Emotion recognition is an important topic in Music Information Retrieval (MIR) [23], but it is typically studied with waveform representation of music. Thus, a reasonably large labeled dataset called VGMIDI has been created to support this research. All pieces in the dataset are piano arrangements of video game soundtracks. A custom web tool has been designed to label these piano pieces according to a valence-arousal model of emotion [41]. Labeling music pieces according to emotion is a subjective task. Thus pieces were annotated by 30 annotators via Amazon Mechanical Turk and the average of these annotations is considered the ground truth. Due to its subjectivity, this annotation task is considerably expensive and hence

the VGMIDI is still a limited dataset with only 200 labelled pieces and 3640 unlabelled ones.

Given the limitation of labeled data, the focus of this work is on search methods that use a music emotion classifier to steer the distribution of a LM. With this framing, a LM is trained with the unlabelled data and a music emotion classifier is trained with the labeled data. Both recurrent neural networks and Transformers were considered as LMs. In order to boost the accuracy of the emotion classifier, it is trained with transfer learning by fine-tuning the LM with an additional classification head. Three different search approaches have been explored to control the LM with the emotion classifier: Genetic Algorithms, Beam Search and Monte Carlo Tree Search (MCTS).

Inspired by the work of Radford et al. [34], the first explored approach is a genetic algorithm that optimizes the neurons in the LM that carry sentiment signal when fine-tuned with a classification head. In this first work, only the valence (sentiment) dimension of the labeled pieces was used to simplify the problem. The LM is a Long short-term memory (LSTM) neural network trained with the 728 unlabelled pieces of the first version of the VGMIDI dataset. The emotion classifier is a single linear layer stacked on top of LM, which is fine-tuned with the 95 labeled pieces of the first version of the VGMIDI dataset. This approach was evaluated with a listening test where annotators labeled three pieces generated to be positive and three pieces generated to be negative. Results showed that the annotators agree that pieces generated to be positive are indeed positive. However, pieces generated to be negative are a little ambiguous according to the annotators.

The second approach is a variation of beam search called Stochastic Bi-Objective Beam Search (SBBS). Beam search is one of the most common methods to decode LMs in text

generation tasks such as machine translation. Traditionally, Beam Search searches for sentences that maximize the probability as given by the LM. Aligning with the work of Holtzman et al. [19], SBBS guides the generative process towards a given emotion by multiplying the probabilities of the LM with the probabilities of two binary classifiers (one for valence and one for arousal). At every decoding step, SBBS samples the next beam from this resulting distribution. SBBS applies *top k* filtering when expanding the search space in order to control the quality of the generated pieces. SBBS was evaluated in the context of tabletop role-playing games. A system called Bardo Composer was built with SBBS to generate background music for game sessions of the Dungeons & Dragons game. A user study showed that human subjects correctly identified the emotion of the generated music pieces as accurately as they were able to identify the emotion of pieces composed by humans.

SBBS is a relatively fast way to decode LMs to generate music. However, it can generate much repetitive music since that maximizes both the probabilities of the language model and the emotion model. In order to improve the quality of the generated music, the third method is a Monte Carlo Tree Search (MCTS) that, at every step of the decoding process, use Predictor Upper Confidence for Trees (PUCT) to search for sequences that maximize the average values of emotion given by the emotion classifier. MCTS samples from the distribution of node visits created during the search to decode the next token. Two listening tests were performed to evaluate this method. The first one evaluates the quality of generated pieces and the second one evaluates the MCTS accuracy in generating pieces with a given emotion. An expressivity analysis of the generated pieces was also performed to show the music features being used to convey each emotion. Results showed that MCTS is as good as SBBS in controlling emotions

while improving music quality.

This dissertation has multiple contributions. The first one is the VGMIDI dataset, one of the first datasets of symbolic music labeled according to emotion. Framing the problem of music generation with controllable emotion as steering the distribution of music LMs with an emotion classifier is also a contribution. The three search methods explored are also significant contributions of this dissertation, once they present the first approaches to solve the proposed problem. Finally, the system Bardo Composer system designed to evaluate the SBBS method is another contribution for being the first system to compose music for tabletop role-playing games.

These contributions open new possibilities of research that can be explored in the future. First of all, different neural network architectures, such as VAEs, can be designed to control perceived emotion in music. VAEs have the benefit of learning disentangled representations of music [51], so they can be a good approach to control emotion in music generation. Second, one can use different search methods to steer the distribution of the language model. For example, new MCTS or SBBS algorithm variations can be designed to improve music quality while keeping the desired target emotion. Third, increasing the VGMIDI dataset can considerably improve the accuracy of the music emotion classifier. This can support the design of deeper neural networks that can potentially be better music generators with controllable emotions. Finally, all these contributions allow different applications, such as composing music in real-time for tabletop role-playing games. One can extend Bardo Composer to create soundtracks for other experiences such as spoken poetry, bedtime stories, video games and movies.

Bibliography

[1] International e-piano competition. http://www.piano-e-competition.com. Accessed: 2019-04-12.

[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

[3] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation-a survey. *arXiv preprint arXiv:1709.01620*, 2017.

[4] Sixian Chen, John Bowers, and Abigail Durrant. 'ambient walk': A mobile application for mindful walking with sonification of biophysical data. In *Proceedings of the 2015 British HCI Conference*, British HCI '15, pages 315–315, New York, NY, USA, 2015. ACM.

[5] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[6] Hannah Davis and Saif M Mohammad. Generating music from literature. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)*, pages 1–10, 2014.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[9] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[10] B Eno. Generative music, speech at the imagination conference, 1996.

[11] Andy Farnell. An introduction to procedural audio and its application in computer games. In *Audio mostly conference*, volume 23. Citeseer, 2007.

[12] Lucas N Ferreira and Jim Whitehead. Learning to generate music with sentiment. In *Proceedings of the International Society for Music Information Retrieval Conference*, ISMIR'19, 2019.

[13] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[14] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. Deepbach: a steerable model for bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1362–1371. JMLR. org, 2017.

[15] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[16] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[17] Hermann Hild, Johannes Feulner, and Wolfram Menzel. Harmonet: A neural net for harmonizing chorales in the style of js bach. In *Advances in neural information processing systems*, pages 267–274, 1992.

[18] Lejaren A Hiller Jr and Leonard M Isaacson. Musical composition with a high speed digital computer. In *Audio Engineering Society Convention 9*. Audio Engineering Society, 1957.

[19] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*, 2018.

[20] Andrew Horner and David E Goldberg. *Genetic algorithms and computer-assisted music composition*, volume 51. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 1991.

[21] 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. *arXiv preprint arXiv:1809.04281*, 2018.

[22] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[23] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion

recognition: A state of the art review. In *Proc. ISMIR*, volume 86, pages 937–952. Citeseer, 2010.

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[25] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3499–3508, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[26] Ben Krause, Iain Murray, Steve Renals, and Liang Lu. Multiplicative LSTM for sequence modelling. *ICLR Workshop track*, 2017.

[27] Fred Lerdahl and Ray S Jackendoff. *A Generative Theory of Tonal Music, reissue, with a new preface*. MIT press, 1996.

[28] Eduardo R Miranda, Wendy L Magee, John J Wilson, Joel Eaton, and Ramaswamy Palaniappan. Brain-computer music interfacing (bcmi): from basic research to the real world of special needs. *Music & Medicine*, 3(3):134–140, 2011.

[29] Kristine Monteith, Tony R Martinez, and Dan Ventura. Automatic generation of music for inducing emotive response. In *International Conference on Computational Creativity*, pages 140–149, 2010.

[30] Gerhard Nierhaus. *Algorithmic composition: paradigms of automated music generation*. Springer Science & Business Media, 2009.

[31] Sageev Oore, Ian Simon, Sander Dieleman, and Doug Eck. Learning to create piano performances. In *NIPS 2017 Workshop on Machine Learning for Creativity and Design*, 2017.

[32] Rafael Padovani, Lucas N. Ferreira, and Levi H. S. Lelis. Bardo: Emotion-based music recommendation for tabletop role-playing games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2017.

[33] Rafael R Padovani, Lucas N Ferreira, and Levi HS Lelis. Bardo: Emotion-based music recommendation for tabletop role-playing games. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017.

[34] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *Arxiv*, 2018.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[37] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. PhD thesis, Columbia University, 2016.

[38] Curtis Roads and Paul Wieneke. Grammars as representations for music. *Computer Music Journal*, pages 48–55, 1979.

[39] Adam Roberts, Jesse Engel, and Douglas Eck, editors. *Hierarchical Variational Autoencoders for Music*, 2017.

[40] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.

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

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

[43] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, pages 1–6, New York, NY, USA, 2013. ACM.

[44] Mark J Steedman. A generative grammar for jazz chord sequences. *Music Perception*, 2(1):52–77, 1984.

[45] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[46] Duncan Williams, Alexis Kirke, Joel Eaton, Eduardo Miranda, Ian Daly, James Hallowell, Etienne Roesch, Faustina Hwang, and Slawomir J Nasuto. Dynamic game soundtrack generation in response to a continuously varying emotional trajectory. In *Audio Engineering Society Conference: 56th International Conference: Audio for Games*. Audio Engineering Society, 2015.

[47] Duncan Williams, Alexis Kirke, Eduardo R Miranda, Etienne Roesch, Ian Daly, and Slawomir Nasuto. Investigating affect in algorithmic composition systems. *Psychology of Music*, 43(6):831–854, 2015.

[48] Duncan AH Williams, Victoria J Hodge, Chia-Yu Wu, et al. On the use of ai for generation of functional music to improve mental health. *Frontiers in Artificial Intelligence*, 2020.

[49] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[50] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

[51] Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia. Deep music analogy via latent representation disentanglement. *arXiv preprint arXiv:1906.03626*, 2019.

[52] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.