# Evaluating Algorithms that Learn how to Compose Music from Scratch

## New long– and short–term evaluation metrics for evaluating to machine learning models which learn to compose music

James Owers

A thesis presented for the degree of
Doctor of Philosophy

Supervised by:
Amos Storkey
Mark Steedman

University of Edinburgh, UK

January 2021

*I, James Owers, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.*

# Abstract

Evaluating whether creative content generated by a computer is 'good,' be it music, images, or text, is unsolved and not even well defined. We identify a property of music which is not modelled well, and propose new evaluation metrics for music generation which can be used to distinguish between real and generated data, and thus be useful for automatic quantitative analysis of generation quality.

We focus on symbolic music because …TODO… This is interesting because …TODO… and it has implications for …TODO…

Finally, we make recommendations for how to make progress with respect to music generation and related tasks.

Acknowledgements

Interdum et malesuada fames ac ante ipsum primis in faucibus. Aliquam congue fermentum ante, semper porta nisl consectetur ut. Duis ornare sit amet dui ac faucibus. Phasellus ullamcorper leo vitae arcu ultricies cursus. Duis tristique lacus eget metus bibendum, at dapibus ante malesuada. In dictum nulla nec porta varius. Fusce et elit eget sapien fringilla maximus in sit amet dui.

Mauris eget blandit nisi, faucibus imperdiet odio. Suspendisse blandit dolor sed tellus venenatis, venenatis fringilla turpis pretium. Donec pharetra arcu vitae euismod tincidunt. Morbi ut turpis volutpat, ultrices felis non, finibus justo. Proin convallis accumsan sem ac vulputate. Sed rhoncus ipsum eu urna placerat, sed rhoncus erat facilisis. Praesent vitae vestibulum dui. Proin interdum tellus ac velit varius, sed finibus turpis placerat.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **API** | **A**pplication **P**rogramming **I**nterface |
| **JSON** | **J**ava**S**cript **O**bject **N**otation |
| **MDTK** | **M**idi **D**egradation **T**ool**k**it |
| **SOTA** | **S**tate **o**f **t**he **A**rt |

# Chapter 1

# Introduction

## 1.1 Historical background

...TODO... Give historical background

### 1.1.1 First instance of music generation

...TODO... From Section 1.2 (Briot et al. 2019)

> The first music generated by computer appeared in 1957. It was a 17 seconds long melody named “The Silver Scale” by its author Newman Guttman and was generated by a software for sound synthesis named Music I, developed by Mathews at Bell Laboratories

### 1.1.2 Mozart using mechanical aids for idea generation

...TODO... From footnote 7 in Section 1.2 (Briot et al. 2019)

> One of the first documented case of stochastic music, long before computers, is the Musikalisches Wurfelspiel (Dice Music) by Wolfgang Amadeus Mozart. It was designed for using dice to generate music by concatenating randomly selected predefined music segments composed in a given style (Austrian waltz in a given key).

### 1.1.3 Ada lovelace noting computers could generate music

...TODO... From (Hollings et al. 2018) Ada Lovelace, "Sketch of the Analytical Engine invented by Charles Babbage, Esq., by L. F. Menabrea," Scientific Memoirs, vol. 3, ed. Richard Taylor, 1843, pp. 666-731 (this quote on p 694)

> "Note G" is the culmination of Lovelace's paper, following many pages of detailed explanation of the operation of the Engine and the cards, and of the notation of the tables. The paper shows Lovelace's obsessive attention to mathematical details - it also shows her imagination in thinking about the bigger picture.

> Lovelace overseed a fundamental principle of the machine, that

> the operations, defined by the cards, are separate from the data and the results. She observed that the machine might act upon things other than numbers, if those things satisfied mathematical rules.
>
> Supposing that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent.

Lovelace also has the Lovelace Test of Creativity attributed to her - see (Ariza 2009).

## 1.2 Modern interest and achievements

...TODO...

- Imogen Heap: How AI is helping to push music creativity
- This is the AI Song Contest In the AI Song Contest teams of musicians, artists, scientists and developers take on the challenge of creating a new Eurovision-like hit with the help of artificial intelligence.
- Swooshes, Seaboards, Synths and Spawn
- David Rosen and Scott Miles on the Neuroscience of Music and Creativity
- AI Music Generation Challenge 2020 - (Sturm 2020)

## 1.3 What are algorithms that learn

...TODO... define/introduce machine learning

## 1.4 What is composing music

...TODO...

## 1.5 What does it mean to compose from scratch

...TODO... what is the minimum information we supply as a starting point? What feedback do we give?

## 1.6 Motivation for this work

...TODO...

- why are we focussing on metrics and not human evaluation
- why do we care about 'from scratch'

## Scope of this work ...TODO... Symbolic music only

## 1.7 List of contributions in this thesis

...TODO...

- [] We review the challenges addressed and models presented amounting to the current state-of-the-art with respect to algorithmic music composition

# Chapter 2

# Literature Review

...TODO...

- ☐ Add content and TODOs from Quip:
  - ☐ https://quip.com/vVjVADMamfDm/2-Literature-Review
  - ☐ https://quip.com/v0MOAQGvMH3O/Literature-Review-Org-Notes
  - ☐ move data and notes tables in ./tables (use YAML, allows large text blocks for notes easy to convert to table with python)
- ☐ outline the different problems people currently try to / can solve, and how these problems relate to 'being able to compose'
- ☐ Review available metrics
  - ☐ Motivate the need for automated evaluation metrics
  - ☐ Why has there been more work on audio than symbolic?
  - ☐ Motivate the need for better evaluation for both short and long term by highlighting shortcomings for each method reviewed

- ☐ Describe state-of-the-art generative models for music composition
- ☐ Identify a gap with respect to modelling long term dependencies by outlining claims and proof of them thus far - this is a specific thing we are going show is poorly evaluated
- ☐ Inform the reader about the multitude of different ways we can represent music and their relative strengths and weaknesses
- ☐ Address ethical shortcomings with respect to learning to compose
- ☐ Update bibtex references to non-arxiv reference if available

## 2.1 Challenges addressed in the literature and how they are evaluated

## 2.2 A summary of evaluation methods for creative models

...TODO... how to evaluate generative models with a focus on music - how do people evaluate their success

### 2.2.1 The need for automated metrics

- ☐ Humans are expensive - show some efforts
- ☐ Humans do not agree - show some research proving this
- ☐ Humans are susceptible to change their opinion depending on context - show some research proving this

- ☐ Consistency is key when tracking performance over the long term
- ☐ Attempts to unify metrics and human opinion - give WMT as an example (Anon 2020)

### 2.2.2 Differences between evaluating audio and symbolic outputs

### 2.2.3 The impossible task of satisfying all evaluation requirements with a single metric

- ☐ Should tie a metric with performance for an intended task (Theis et al. 2016)

### 2.2.4 Evaluation metrics and representations used for extracting musical structures

## 2.3 Models for composing music

...TODO... State caveats about our distinctions: 1. Learning is essentially copying 2. By specifying the method of learning, we are incorporating expert knowledge 3. All models must work with a human composer to some extent - the programmer must choose a representation for the music and is therefore a composer in some senses

- ☐ Make and curate comparison table of models:

    - ☐ keep as csv
    - ☐ at min, we can use pandas to read and auto convert for insertion here
    - ☐ is there a way to (**reference?**) the file here and have pandoc insert? <- **do not spend time on this, cursory google!**
- ☐ Music Transformer [@]
- [ ]
- ☐ Extend table from Chapter 7 and information from Chapter 6 in review paper (Briot et al. 2019)
- ☐ Go through https://paperswithcode.com/task/music-generation

### 2.3.1 Models which learn from scratch

…TODO… These are the models which our research pertains to

### 2.3.2 Models which do not learn from scratch

…TODO… These models are stated to highlight why they are different, have an unfair advantage in certain contexts, or explain why they are out of scope with respect to the investigation of this thesis.

#### 2.3.2.1 Heuristic models which primarily copy and edit music from a database

#### 2.3.2.2 Models which incorporate expert knowledge into their design

...TODO... e.g. with respect to structural hierarchy

#### 2.3.2.3 Models which only work in conjunction with a human composer

#### 2.3.2.4 Proprietary models: models for which adequate details of their design are not publicly available

## 2.4 Methods for representing music on a computer

...TODO... how to represent music data - (in relation to 'from scratch,' what is the minimal information supplied to the models, and is there evidence of what difference it makes (either by experiment or just by reasoning?)

- ☐ Note our desires with respect to our modelling challenges: we want the input to be *minimal and flexible* - the model should learn as much as possible as if it were a human listener
  - ☐ Ideally we would work directly on sound, but this involves an

additional layer of representation.

### 2.4.1 Information that must be captured about a musical performance

### 2.4.2 The different representations of symbolic musical information

#### 2.4.2.1 Summary of differences

- ☐ Highlight where *information* captured by each representation is both **different** and **more/less amenable to being learned**

### 2.4.3 Availability of data for each representation

- ☐ Quantity,
- ☐ Quality,
- ☐ Legal issues

### 2.4.4 Availability of software for different representations

- ☐ Describe MusPy (Dong et al. 2020) for conversion between data formats
- ☐ Describe Music21 (Cuthbert & Ariza 2010) for conversion between data formats

### 2.4.5 Evidence from the literature regarding modelling performance differences

- ☐ Find any reviews (or lack thereof) of model performance differences with respect to:
  - ☐ evaluation metrics
  - ☐ speed

## 2.5 Ethical considerations when designing automated methods for composing music

# Chapter 3

# New metrics for Evaluating Musical Generations

## 3.1 The midi degradation toolkit

## 3.2 A phrase-level metric for short-term structure

## 3.3 A piece-level metric for long-term structure

# Chapter 4

# Evaluating State-of-the-Art Music-generating Models

## 4.1 Comparative analysis using new and existing metrics

- Use phrase and piece level metrics to evaluate state-of-the-art models
- Compare and contrast, outlining the issues identified (e.g. meandering, no high-level structure)

## 4.2 Strengths and shortcomings of existing models

## 4.3 Avenues for improvement

# Chapter 5

# A New Model

Potential ideas:

- An improved generative model for music
  - Training like BERT? http://jalammar.github.io/illustrated-bert/
  - Using mdtk for data augmentation in training (negative examples?), making them more robust
  - Alternative training objectives:
    * crossentropy slow and not musically informed
    * can we use something akin to word error rate (this has been done for text)
- Alternative ways to encode music: encoding chords and phrases in a low-rank continuous space
  - Have done some work on this with convnets and generating continuations
    * low rank was enforced by cross-product ing two vecs

- Could investigate effect of different representations for music on performance

# Chapter 6

# Conclusion

# Chapter 7

# References

…TODO… check over using https://www.cl.cam.ac.uk/~ga384/bibfix.html

Anon, 2020. 2020 Fifth Conference on Machine Translation (WMT20). Available at: http://www.statmt.org/wmt20/ [Accessed January 23, 2021].

Ariza, C., 2009. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*, 33(2), pp.48–70. Available at: https://www.jstor.org/stable/40301027 [Accessed January 22, 2021].

Briot, J.-P., Hadjeres, G. & Pachet, F.-D., 2019. Deep Learning Techniques for Music Generation – A Survey. *arXiv:1709.01620 [cs]*. Available at: http://arxiv.org/abs/1709.01620 [Accessed January 22, 2021].

Cuthbert, M.S. & Ariza, C., 2010. music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. *Michael Cuthbert*. Available at: https://dspace.mit.edu/handle/1721.1/84963 [Accessed January 23, 2021].

Dong, H.-W. et al., 2020. MusPy: A Toolkit for Symbolic Music Generation. *arXiv:2008.01951 [cs, eess, stat]*. Available at: http://arxiv.org/abs/2008.01951 [Accessed January 23, 2021].

Hollings, C., Martin, U. & Rice, A.C., 2018. *Ada Lovelace: The making of a computer scientist*, Oxford: Bodleian Library.

Sturm, B., 2020. The 2020 Joint Conference on AI Music Creativity. *The 2020 Joint Conference on AI Music Creativity*. Available at: https://boblsturm.github.io/aimusic2020/ [Accessed January 22, 2021].

Theis, L., Oord, A. van den & Bethge, M., 2016. A note on the evaluation of generative models. *arXiv:1511.01844 [cs, stat]*. Available at: http://arxiv.org/abs/1511.01844 [Accessed January 23, 2021].