# Title of your thesis

Your Name

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Your Department

Your Advisor, Chair
First Committee
Second Committee
Third Committee
Last Committee

December 4, 2020
Blacksburg, Virginia

Keywords: Some Keywords, Subject matter, etc.

# Title of your thesis

Your Name

## ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

*Dedicated to Virginia Tech.*

# Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In 1952 L.A. Hiller and L.M Issacson ushered forth a new era of the study of both music and computer science when they introduced the Illiac Suite – the first composition that was created solely by a computer [? ]. What we'll refer to broadly as Computer Music (CM) research has continued to see impressive advancements since the introduction of the Illiac Suite in several different domains, including musical composition[? ], instrument and sound synthesis[? ], and musical analysis[? ]. CM research presents a unique challenge to both musicology and computer science due to the highly subjective nature of music paired with the strong quantitative and mathematical nature of computer science. However, music is also inherently mathematical and contains a strong hierarchical structure from which powerful patterns emerge - how it is that these common patterns lead to such a highly subjective human experience is outside of the scope of this work. It is perhaps due to the inherently paradoxical nature of music that it creates such an interesting set of problems to study, particularly from a computational perspective. In the authors opinion, this problem set is one of the most worthwhile to study in the current day, and should receive more focus in the research literature.

To reach such a point, it is necessary to view the field from the lens of Artificially Intelligent musical systems that are able to reason themselves about music. In general, Artificially Intelligent systems have seen immense progress in the last decade due to the rise of Machine Learning (particularly with Deep Learning) and it's applications in several different domains.

Music has been one of these domains and has seen impressive advances in several musical tasks such as musical composition[? ], and musical analysis[? ].

One of the more intriguing problems in computer music is the creation of an expressive performance generation system. There are several commercially available notation and playback software systems [1] that are able to automatically generate musical performances from a purely symbolic musical representation in the form of a score (more commonly known as sheet music). The systems are built based on a predefined set of rules that create deterministic performances given a score. Although the performances are technically an ”accurate” rendering of the score, they don't contain the *human* element. That is, they contain a straight and deterministic mapping between a note marked in a score and its corresponding position in the associated performance. In such systems, there is no *expression* of the performance. As such, these systems produce robotic-sounding performances that are offputting to human listeners [2].

These simple performance generation systems do not render ”expressive” because they don't account for performance features that add the human element. Such features include variations in tempo, timing, and dynamics (TODO: Add a reference to a later section that will go over each feature in detail). A performer of a composition uses each of these features of the performance to add a unique interpretation of the piece. This interpretation (or expression) is responsible for providing the ”human” element of musical performance.

This poses the question of whether or not a computer system can create expressive performances that are close to actual human performance (or even creating a completely new style of expression that a human cannot). This has been an active area of computer music research since the 1980s [? ]. This thesis is a further exploration of the current research

[1] `musescore.com` and `www.finalemusic.com`

[2] an example performance can be heard from musescore

with a particular emphasis on applying the current state of the art Deep Neural Network architectures from Natural Language Processing (NLP) research to the problem domain and understanding their effects.

Richard

Revisit introduction after outline and more writing

# Chapter 2

# Background

The background section needs to include information relevant research and clearly set up a research problem.

## 2.1 Expressive Performance Generation

Define expressive performance generation (EPG) at a technical level (data features). Give background into how it fits into MIR research.

- Define a score and Performance
  - Explain possible features for both score and performance
  - Richard: Create (find) figures for score and performance
- Explain how expressive performance generation fits into music generation research
  - Generation as subset of MIR research
  - Richard: Create graph showing (or referencing other graphs) of where performance generation fits into music generation as a whole
- Papers to cite

  - This time with feeling [? ]
  - Deep learning for music generation survey [? ]

## 2.2 Transformers

Provide context to why transformers are important and the problems they've solved in nlp.

- Intuition behind transformers and why they are so powerful in sequence modeling
- Attention is all you need paper [? ]
  - State of the art in translation tasks
  - New architecture for sequence modeling using only attention. No recurrent network
- BERT [? ]
  - Transformer Encoder only
  - Self-supervised learning and pre-training. Includes having a simple multi-layer perceptron at the end to make it useful
- Music Transformer [? ]
  - Builds off of This Time with Feeling[? ] paper. Both composition and performance generation at the same time
  - Implements full transformer architecture
  - Achieves better results than LSTM

- Question: Can a transformer model be applied to only performance generation with an encoder only architecture to achieve better results than current state of the art models?. Intuition says yes given the results from Music Transformer.

## 2.3 Existing EPG models

- KTH system [? ].
- YQX. A bayesian network that models timing, dynamics, and articulation [? ]. Won the 2008 RenCon contest.
- Basis Function Models [? ] (charcon phd thesis)
  - Linear Basis Functions. Uses Least Squares regression and Bayesian models with about the same performance
  - Non-Linear Basis Functions. Uses standard feed forward network. FFNN perform better than Linear models. Also uses an RNN.

Richard

This needs more exploration. Lot of possibilities for future work

- Giraldo and Ramirez use several different ML algorithms, including Decision Trees, k-NN, SVM's, and FFNN to build an expressive performance generation system for improvisational Jazz guitar [? ].
- Moulieras and Pachet use a Maximum Entropy model to infer the underlying distribution of expressive performance and build a generation system trained from a mix of popular music. Their expressive model outperforms base models in listening tests [? ].
- Jeong builds two versions of virtuosoNet, one using a recurrent hierarchical attention network (HAN) [? ], and another using a recurrent graph network [? ]. These models are built using a dataset order of magnitudes larger than other datasets and attempt

to model the expressive performance feature of the pedal, which no other model does. The code for the models is also open source so it was chosen as the starting place for this work.

Richard: Add more papers and expand upon the existing research a bit more. Isn't completely necessary but will be good for my overall understanding

## 2.4 Datasets

- Talk about the fundamental limitations of gathering data for this problem, especially in relation to other fields [? ]. Because of this, the lack of high quality data is limited.

- The dataset used for the virtuosoNet [? ] [? ] will be the dataset used for the experiments. At the time the experiment started it was the largest publicly available dataset applicable to the EPG systems, and was chosen for use. A recent publication [? ] builds off of the dataset used for the virtuosoNet with more sophisticated alignment and some extensions to the size (dataset is named ASAP). ASAP would be more appropriate for future use.

- One of the necessary data processing tasks for EPG is the alignment between the score and performance of a given piece. Because there is always an inherent interpretation of a composition by a performer, there is no clear mapping between any given score and performance. It remains necessary to have some sort of alignment process to match each note in the performance with it's related position in the score.

  Richard: Reference this in the introduction

  Richard: This needs more research. Find relevant papers to cite, as well as show a diagram that makes it clear why alignment is necessary

.

## 2.5 Evaluation

- Evaluation is particularly difficult for a problem like EPG because there is no ”correct” interpretation of a score. However, there is at least a vaguely understood relationship between a score marking and how a performaner should use that marking within the context of a performance. For example, if a crescendo marking is used in a score, the performer should at the very least increase the volume of the performance relative to the current volume of the piece. The amount which the volume should increase or the rate at which it increases are not clearly defined, but the fact of the increase of volume itself is. This is the fundamental intuition behind the motivation to build computational models for expressive performance. Nonetheless, it still remains a difficult job to evaluate a given EPG model because of the ambiguity of what is ”correct” or not.

- Evaluation methods used so far in EPG models are broken into two categories, quantitative and qualitative.

- Quantitative:

  - This follows standard techniques for experimentation of evaluation of ML models in general. It usually involves calculating a numerical value for a models inference on a separate test data set that was not used for model training or model selection. . Common metrics for regression like problems are mean squared error (MSE) and the pearson correlation coefficient (R2).

Richard

Find reference for ML training and evaluation

  - Due to the nature of EPG model evaluation mentioned above, it is not clear that ”better” quantiative metric score for a given model over another indicates that the performance of the model is superior. .

Richard

Find section in Garcon survey that references this point

- Qualitative
  - Qualitative evaluation methods involve gathering human feedback by playing performances of a given models performance to an audience and getting ratings or judgement of the model according to a predefined questionnare or survey method. The nature of these evaluation methods is not consistent in the current literature and remains a challenge for the field to solve in the future. .

    Richard

    Find section in Garcon survey that references this point
  - Richard: Conduct more research for reference on current methods for qualititative evaluation

# Chapter 3

# Methods

## 3.1 Model and Experiments

- Due to the open-source nature of virtuosoNet project and its attempt to build a more cohesive EPG model by introducing the pedal as an expressive feature and training on a much larger dataset, we built off of this model.

- Because of the significant advances in other sequence modeling domains (such as NLP) and the indication of increased performance of another related task with the Music Transformer [**?** ], the main question we want to answer is whether we can see similar increases in model performance by applying a Transformer ANN architecture to the problem domain.

- We will experiment with a transformer encoder only architecture similar to BERT. The problem includes a 1-1 to mapping between every note in the score and a related note in a performance. This is different than seq-2-seq modeling problem such as neural machine translation which maps a sequence of one length to another sequence of a different length, which is what the full Transformer architecture was intended for. The Transformer Encoder can be seen as as a large encoder that learns the best represenation for a given feature set. The model we'll build will use a simple FFNN that accepts the output of the transformer encoder to decode this representation and give the final feature set which is then used to create a performance. This is similar to the BERT architecture and it's intended application.

Richard: Come up with more detailed explanation of this modeling choice. Also create a visual diagram that explains the transformer encoder with the simple regression model sitting on top of it

- Because we are using the same dataset used to train virtuosoNet, we will directly compare the performance a Transformer model to the existing virtuosoNet models using the same quantitative metric, MSE.

  Richard: Come up with specific model experiments and comparison in a table. Table doesn't have to have results but needs the general outline that will be used in the final paper

## 3.2 Evaluation

- Quantitative: Because we are using the same dataset used to train virtuosoNet, we will directly compare the performance a Transformer model to the existing virtuosoNet models using the same quantitative metric, MSE.

  Richard: Come up with specific model experiments and comparison in a table. Table doesn't have to have results but needs the general outline that will be used in the final paper

-

  Richard: Come up with qualitative evaluation framework

# Chapter 4

# Results

## 4.1 Quantitative

Add table with results of experiments along with explanations.

## 4.2 Qualitative

Add graphic with results of experiments as well as explanations.

# Chapter 5

# Discussion

Richard: The following are some interesting discussion ideas that have come up so far. There is no telling if these will be in the final paper or not after conducting more experiments.

- Transformer performs worse according the quantitative metrics. This could be because it doesn't build in a specific hierarchical layer that is specific the problem. It is a much more generic model. There is a lot of room for exploration into experimenting with different architectures based on the Transformer to better fit the problem domain.

  Richard: Add more discussion based on more results

- Transformer appears to be a more dynamic model than the recurrent virtuosoNet model that makes more "mistakes". Does this mean that it is more "human".

  Richard
  Add discussion of uncanny valley

- Pedal in performance is messy. Could be because of problems in the feature and modeling, or could just be because it is a difficult problem to model.

-

  Richard: Discussion on qualitative results

# Chapter 6

# Conclusion

- Re-iterate high-level results. Transformer performs worse according to a quantitative metric. Doesn't necessarily mean that it's a "worse" model
- Give a more in-depth discussion of the challenges of the problem domain with added context given the technical details discussed in our paper.
  - Large need for more high-quality data. There are efforts on this front [**?** ] - our model could easily apply to this dataset. I have the intuition that the alignment and consistency in the score data extremely important to building more robust and general models. The dataset used is large which adds it's own advantage, but it inherently presents more room for error.
  - Better methods for evaluation. This is a current pen question in EPG research, and could constitute an entire area of study outside of build the EPG models. Better evaluation methods would provide more intuition on how the Transformer model performs in comparison with other models.
  - Modeling of pedal appears to be a hard problem to solve, and hasn't been done outside of virtuosoNet and this work.
- Future research directions
  - Experiment with different transformer architectures. Immediate experiment to run is mess with the positional encodings as was done in [**?** ]

  - Build and find better data
  - Explore evaluation methods
  - See how Transformer model fits into other EPG frameworks and datasets, first example being the BM framework proposed by [? ]. Also would be useful to apply it to other genres besides solo classical piano

- Research Applications
  - Tutor systems
  - It's place in a full end to end music generation system.
  - Direct application to notation software and the creative process of composers.

- Philosophical questions about problem domain: How can a EPG can help to improve our understanding of music itself. If we can learn to generate human-like musical performances, does that mean we understand what it is that constitutes the "human" element of music? Another even more philosophical question would be - can we understand what it means to "human" in general. Music is one of, if not the most, unique human experiences. If building an EPG model can't directly answer the question of what it means to be human, it can at least provide insight.

# Appendices

# Appendix A

# Appendices I

## A.1 A1

## A.2 A2