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 vii

List of Tables viii

1 Introduction 1

2 Background 2

2.1 Expressive Musical Performance 2

2.1.1 Score and Performance 4

2.2 Data 7

2.3 Transformers 8

2.4 Evaluation 9

3 Related Work 11

3.1 Existing EPG models 11

3.2 Datasets 12

4 Experiments 14

4.1 Model and Experiments 14

4.2 Evaluation 15

5 Results **17**

5.1 Quantitative 17

5.2 Qualitative 17

6 Discussion **18**

Appendices **19**

Appendix A Appendices I **20**

A.1 Musical Concepts and Terminology 20

A.1.1 Pitch 20

A.1.2 Tempo and Timing 20

A.1.3 Dynamics 21

A.2 A2 21

List of Figures

List of Tables

Chapter 1

Introduction

- Introduce the idea of musical research with computers. Talk about the illiac suite [**?**] and Music Information Retrieval.
- Significance of machine learning on the field
- Introduce idea of expressive musical performance. Brief conversation about the different performance components (articulation, dynamics, timing).
- Using Transformer architecture which hasn't been done in the field.
- Richard: report results

Chapter 2

Background

There are two major research components that this project is based on. The first is the problem domain of expressive musical performance (EMP), and the second is the ML modeling domain of Transformers. We will introduce both of these components and provide context for what makes them interesting as a research project and why they are worth exploring together. We start first with an overview and definition of EMP, and then a summary of the Transformer.

2.1 Expressive Musical Performance

EMP is a subset of the research field of Music Information Retrieval (MIR) [1] whose purpose is to use computational information to study, interpret, and gain a better understanding of the *essence* of music itself. One of the most important components of music is how it is performed. Musical performance facilitates the communication of a musical composition created by an entity (most commonly a human) and the auditory and cognitive experience that is perceived by an audience. Each componet of the musical process can be viewed as a role rather than a person, as it is often the case that the composer, performer (and even the listener), are all the same person. Even if this is the case, each role in the musical process

[1] Widmer[**?**] points out that MIR itself does not encompass the entire scope of computer music research, but that it is a good proxy to use when referring the field as a whole. We will operate under the same assumption

is unique in it's responsibility and musical experience. We are interested in the role of the performer - and more specifically, how the performer uses *expression* to increase the quality of the musical experience.

Perhaps the most important role of a musical performer is to add their own unique interpretation of a musical composition. This is especially important in the case of western classical music, wherein the performer(s) are encouraged to use their own musical knowledge and skill to add something of their own to the musical piece. This is due in part to several things. The first is the nature of the way the music is composed in the first place, given that the composer often intentionally leaves the aspects of the performance up the interpretation of the performer. The second is the inherently dynamic nature of the musical instruments used to generate the performance. They are created in such a way that the same piece of music, whether it be an entire composition, a musical phrase, or even a single note, can be played in almost an infinite amount of ways . Not only would the exact reproduction of a musical performance not be desirable from the listeners perspective, it would also be almost impossible to achieve from a performers perspective. This creates an inherent interpretation or expression in every performance of a composition.

Richard

Find reference for differences in same performer and same piece

There are a number of interesting research questions that arise when considering EMP from a computational perspective. They can be broadly categorized in two ways - the first as methods for EMP analysis, the second as methods for EMP generation [**?**]. We are interested in the latter. More specifically, we are interested computational models that are used to generate musical performance, as opposed to models that are used to better the understanding of musical performance by analysing existing performances. [2]. In order to a build a model that can generate a musical performance, it is nessary to create a detailed

[2]Of course the two problems are not unrelated and the advancement in one implies advancement in the other. However, it is still useful to conceptualize them as separate research questions that have a large overlap.

abstract definition of EMP that the model can implement. Our focus will be on solo piano western classical music

2.1.1 Score and Performance

EMP usually involves two different components; a musical score which is a symbolic representation of a musical composition, and a musical performance which encodes different expressive parameters related to the performance. The most traditional interface that represents a score is sheet music, which exists in both physical and digitial form and is a graphical representation of a musical composition. Historically, the only way to represent a musical performance has been directly in audio generated by acoustic instruments. However, the advent of digital music has given rise to electronic synthesizers and data formats which allow for more diverse representations of musical performance. The most commonly used data format is MIDI, which is a file format and data transfer protocol which encodes different elements of a musical performance into an event based representation.

Before we dive into the details on the differences and similarities between a score and a performance, it is important that the reader has a sufficient level of musical structure and terminology. There are several key concepts that, although they may appear elementary, are necessary to understand and define in such a way that they can be useful when used with a computational model. These concepts are outlined in appendix A.1

Score

A musical score contains all of the information that relates to a musical composition, and can be thought of as a higherical structure that presents information about the composition at different levels. The lowest level contains information about the pitch and timing of every single note, as well as optional information about how the note should be played. This can include information specific to instruments such as the bow direction of a violin, but for our purposes (dealing only with piano) we will consider this to be the articulation of each note, usually indicated by legato or staccato

Richard: Make sure to have some background information on articulation in the appendix

The middle level contains information related to certain substructures within the musical composition, which are usually expressed within a grouping of notes or measures. The most common score annotations at this level are dynamic markings which indicate whether to play a grouping of notes loud (Forte), soft (Piano), or to gradually increase or decrease the volume (crescendo or decrescendo). Although dynamic markings are the most common at this level, it is also possible to see score markings for all other musical features, such as local tempo or articulation of a certain substructure. Perhaps the most important score marking at this level is that of a phrase, which is a marking that indicates that a group of notes should be interpreted as belonging to a singular musical idea and that each note should fit within the context of the phrase as a whole. A phrase can be expressed through all of the different aforementioned musical features, including the tempo, timing, dynamics, and articulation of the notes.

The highest level contains meta information that relates to the entire composition as a whole. This information typically includes the key signature and time signature, as well as the global tempo for the entire piece, most commonly represented as BPM.

Peformance

An expressive musical performance contains most of the same musical information as does a score, but with one key difference; that is, that an expressive performance will deviate (or interpret) from the exact information that is presented in the score. For example, although a score may indicate a tempo of 120 BPM, it is highly unlikely that a given performer will perfectly adhere to this tempo throughoug the entirety of the piece. This is even more apparent if the score indicates a change in tempo somewhere in the composition. If a score indicates that the performance should speed up over a series of notes, there is no telling at what rate the tempo should increase. Some performers may choose to speed up at a fast rate and over a short period of time. Others may choose to increase the tempo at a slow rate and over a longer period of time. A single accelerando (a score indication to pick up the tempo) can result in either of these outcomes.

With that being said, a performance contains most of the same features related to a score, which include pitch, tempo, timing and articulation. Each of these expressive features will be measurable and absolute, whereas the score markings of these features can be viewed more as a suggestion than a rule. There a few additional features that are present in performances which are not in scores. The first we will refer to as deviation which is heavily related to timing. It can be thought of as a numberical number which represents how far off the timing of a particular note deviates from it's "correct" position in the score. These micro-timing deviations present in musical performances are an essential part of expression. Without them, indicating that each note onset and offset is exactly in line with the it's marking in the score, performances sound robotic and mundane .

Richard

add reference and sample performance

The other important feature of performance that is not always present in a score applies specifically to the piano, and is the presence of a piano pedal. There are several different types of piano pedals, but the most common are the sustain pedal, which prolongs the duration of every note of the piano when activated, and the soft pedal which softens the sound of the entire piano. Although the effects of these pedals are directly related to the articulation and dynamics of the performance, their presence (or lack of) can be seen as a crucial component of piano performance. It is common for the sustain pedal to see active use in almost all modern piano performance, even when there doesn't exist any score marking indicating it's use.

Richard: Add section and reference to the specifics of feature engineering related to both the score and the performance in the methods section

.

2.2 Data

A brief section about the data used for the problem. Introduce MusicXML and MIDI

- MusicXML
  - A text based representation of a musical score.
  - Created as a way to standardize score data among different notation software.
  - Useful for EMP research because of the standardized format.
  - Contains all relevant information about the score and it's related features.

Richard

Add reference to feature section

- MIDI
  - Event based protocol for digital representation of musical instruments.
  - Used in a variety of ways, most commonly known for it's use in DAW software to represent easily editable tracks for music production.

- Can be synthesized in many different ways.
- Contains all of the needed information to represent a musical performace. .

Richard Reference feature section

2.3 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.

2.4 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

Richard

Find section in Garcon survey that references this point

- 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. .

- Qualitative

Richard

Find section in Garcon survey that references this point

  - 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: Conduct more research for reference on current methods for qualititative evaluation

Chapter 3

Related Work

3.1 Existing EPG models

- KTH system [?]. A rule-based system for expressive performance. Rules are selected through a empirical process based on human feedback.
- YQX. A Bayesian network that models timing, dynamics, and articulation [?]. Won the 2008 RenCon contest.
- Basis Function Models [?]
  - 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

3.2 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 EMP 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 Richard: Reference this in the introduction, 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 its related position in the score.

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

.

Chapter 4

Experiments

4.1 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 representation 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 a 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

4.2 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. The table doesn't have to have results but needs the general outline that will be used in the final paper

- Due to time and resource constraints, no sophisticated qualitative evaluation was conducted for the models. However, a personal evaluation was used during the entire model development process.

Richard: Talk about method used for personal analysis

-

Chapter 5

Results

5.1 Quantitative

Add table with results of experiments along with explanations.

5.2 Qualitative

Give personal qualitative report.

Chapter 6

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

Appendices

Appendix A

Appendices I

A.1 Musical Concepts and Terminology

A.1.1 Pitch

The first and most basic component in music is pitch. Pitch is a perceptual property of sounds that relates to the physical frequency of a sound vibration [?]. It is what determines whether or not a sound can be though of as ”high” or low”. The most commonly known way to conceptualize pitch is the 88 different keys on a piano keyboard, where each key represents a different patch value. Pitch is most commonly labeled using scientific pitch notation, which couples a range of letters (A to G) with a range of numbers (zero to eight) that correspond to different octave ranges [1]. The most well known pitch is C4, or ”middle C”, and lays in the very center of a standard 88 key piano.

Richard

Create or find visualization

A.1.2 Tempo and Timing

Tempo in music describes the rate at which notes are played, and timing describes when a particular note should be played relative to the start of the composition. They are best explained in the context of modern western musical notation introduces the idea of note

[1] https://en.wikipedia.org/wiki/Scientific_pitch_notation

durations, time signatures, measures, and beats [2].

Richard: Find a more intuitive way to explain this. The piano roll explanation and visualization may work better

Each composition is broken down into a sequence of measures, and the time signature defines how many beat exist per measure, as well as the duration of a single beat. For example, a 4/4 time signature indicates that there are 4 beats per measure (the top half of the time signature), and that the duration of each beat is represented by a quarter note. A 3/4 time signature would indicate only 3 beats per measure, with the beat duration represented by a quarter note. The timing of a note would refer to it's measure, beat, and note duration. Tempo is most commonly given in beats per minute (BPM). A composition with a 4/4 signatue and a 120 BPM would mean that after one minute, 30 measures of the composition should have been played so far.

Richard: create or find visualization

A.1.3 Dynamics

Dynamics can simply be thought of as how loud or soft a note should be played (or has been played).

A.2 A2

[2] See https://en.wikipedia.org/wiki/Musical_notation#Modern_staff_notation for a more detailed explanation