Neural Translation of Musical Style

Iman Malik
Department of Computer Science
University of Bristol
Bristol, United Kingdom
im13557@bristol.ac.uk

Carl Henrik Ek
Department of Computer Science
University of Bristol
Bristol, United Kingdom
carlhenrik.ek@bristol.ac.uk

1 Introduction

Music is mysterious. Anthropologists have shown that every record of human culture has some aspect of music involved [1]. However, the exact evolutionary role of music is shrouded in mystery. Scholars theorise and state that music must have emerged as an evolutionary aid [2, 3]. Some propose that the function of music was to provide social cement for group action [2, 4, 5]. War songs, national anthems, and lullabies are all examples of this.

Music is fundamentally a sequence of notes. A composer constructs long sequences of notes which are then performed through an instrument to produce music. However, when music is performed from sheet music, it needs to be interpreted. The ambiguity in interpretation results in a variety of different realisations of the same sheet description. In abstract terms, this means that the mapping between the sheet notation and the performed music is not a bijection.

This leads the central question of this paper, *is it possible to leverage data and learn how to automatically synthesise musical performances that are indistinguishable from a human performance?* Specifically, we postulate that a significant portion of the style injected by a musician comes from **dynamical aspects**. To that end, we aim to learn to inject the note **velocities** from data only containing the note **pitches** over time.

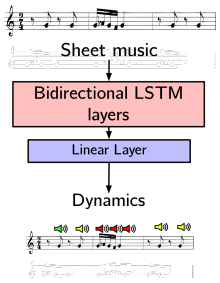

Figure 1: GenreNet architecture.

2 Model

GenreNet

We describe the neural network architecture of "GenreNet". GenreNet predicts the dynamics of a given sequence of notes. The model consists of bidirectional Long Short-term Memory (LSTM) [6] layers followed by a linear layer as shown in Figure 1.

StyleNet

However, GenreNet is limited to learning the dynamics of a specific genre. As stated in the introduction, we would also like to investigate whether it is possible for the model to learn different styles. Thus we propose the network architecture named "StyleNet" which is shown in Figure 2. This model has an LSTM layer which is shared amongst GenreNet units. Each unit is responsible for learning a specific style.

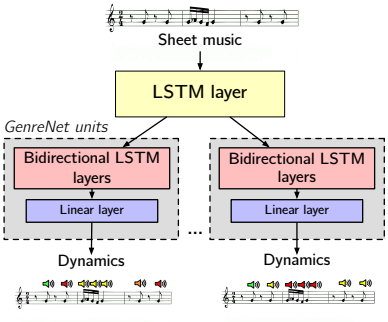

Figure 2: StyleNet architecture.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

3 Training

For training purposes, we first create a human-performed dataset from which StyleNet can learn different musical styles [7]. We present the Piano dataset which contains Piano MIDI files within the Classical and Jazz genre. All MIDI files are in ¼ time, format 0 and are human recordings. Both genres have 349 MIDI files creating a total of 698.

We quantise the dataset and design input/output matrix representations. Please refer to the Appendix for supplementary material. The parameters of StyleNet are optimised by minimising the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 \tag{1}$$

where y_i is the recorded velocities of notes played at time-step i, f_i is the predicted velocities and n is the number of time-steps in the song.

StyleNet was successfully trained on alternating batches of Jazz and Classical music using the ADAM optimiser [8] on a Nvidia GTX 1080 Ti. A dropout [9] of $p = 0.8$ was applied, and gradients were clipped by norm where $g = 10$ with a learning rate of 10^{-3}. The final model was trained for a total of 160 epochs.

4 Experiments

How does one evaluate a musical performance? Music only holds meaning through the confirmation of a human. A decreasing loss shows us that the model is trying to understand the problem numerically. However what one wants is to minimise the "perceptual" loss. Thus it can be quite challenging when trying to evaluating a model in the field of music.

We designed experiments to answer the following questions:

- Can StyleNet's performances pass as a human's?
- Can StyleNet perform the same musical score in different styles?

Identify the Human: To evaluate whether the model can perform like a human, the "Identify the Human" survey was set up. Participants are shown two 10 second clips of the same performance where one is generated and the other is an actual human performance. Participants need to identify the human performances. We use a baseline of randomly guessing between both performances. An average of **53%** from the participant pool could highlight the human performance. This is a surprisingly low number which concludes that StyleNet's performances pass as a human's.

Identify the Style: This leads the next investigation into the model's ability to play a score in a specific style. The "Identify the Style: Classical or Jazz" survey was set up for performances generated in Classical and Jazz styles. Two stylised tracks of a score are shown to the participants. Participants need to correctly identify the style being asked for.

An average of **47.5%** respondents selected the correct style. Similar to the previous test, the baseline of this test is randomly guessing between both answers. The analysis of this number shows that the structure of StyleNet is not sufficient to separate the characteristics of the two styles. We believe that this could be the result of several different factors, for one, we do not have examples of the same sheet interpreted in both styles. Such data would encourage the style split at the shared layer in the model. Furthermore, style is something that is "added" to the composition which might be challenging to capture with this sequential structure.

Identify the Human (Extended Performance): Many participants mentioned that 10 seconds is too short to decide on an answer. It can be hard to assess a short clip without its surrounding musical context. Thus a consolidating experiment for our initial findings would be to assess the model on a complete performance. The experiment set-up was identical to the "Identify the Human" test for short audio clips, but the only difference was that participants had to answer one question featuring an extended performance. The survey was completed by 99 people and **46%** participants could identify the human. This result consolidates that StyleNet can successfully generate performances that are indistinguishable from that of a human.

References

[1] Iain Morley. A multi-disciplinary approach to the origins of music: perspectives from anthropology, archaeology, cognition and behaviour. *Journal of anthropological sciences = Rivista di antropologia : JASS*, 92:147–77, 2014.

[2] Jay Schulkin and Greta B Raglan. The evolution of music and human social capability. *Frontiers in neuroscience*, 8:292, 2014.

[3] David Huron. Science & music: lost in music. *Nature*, 453(7194):456–457, 2008.

[4] Steven Mithen, Iain Morley, Alison Wray, Maggie Tallerman, and Clive Gamble. The Singing Neanderthals: The Origins of Music, Language, Mind and Body. *Cambridge Archaeological Journal*, 16(01):97–112, 2006.

[5] Kevin M. Kniffin, Jubo Yan, Brian Wansink, and William D. Schulze. The sound of cooperation: Musical influences on cooperative behavior, 2016.

[6] Sepp Hochreiter and J Urgen Schmidhuber. LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Yamaha International Piano-e-Competition. *http://www.piano-e-competition.com/*.

[8] Diederik P Kingma and Jimmy Lei Ba. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

A Supplementary Material

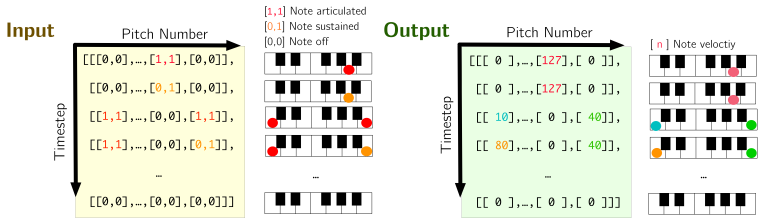

Figure A.3: Input and output representations.

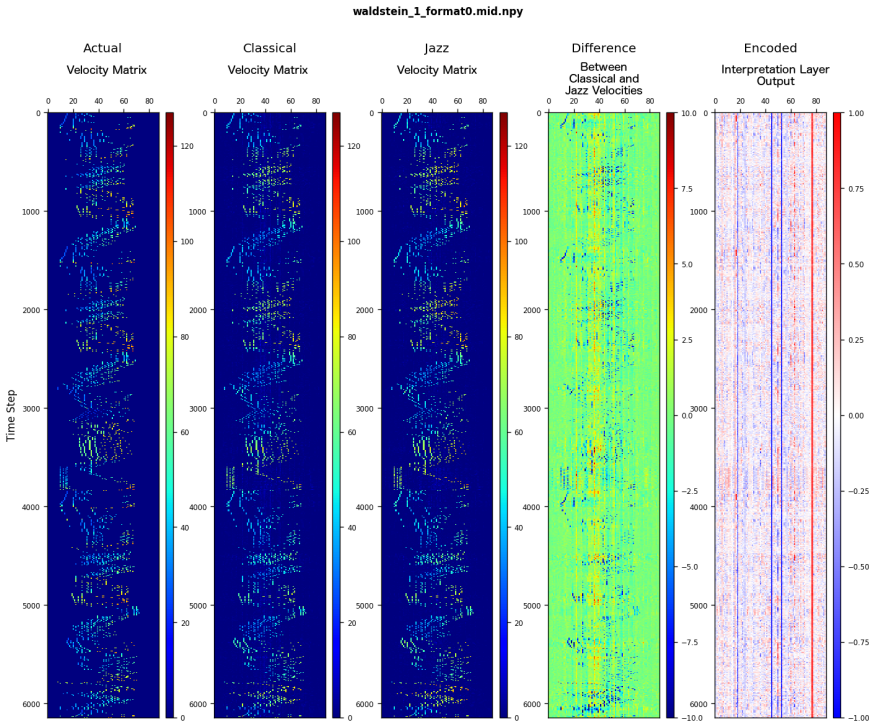

Figure A.4: Training snapshot of StyleNet's predictions for waldstein_1_format0.mid.

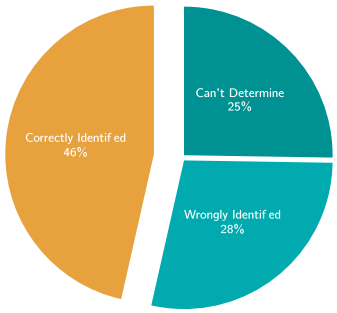

Figure A.5: "Identify the Human: Extended Performance" survey results.