An Analysis of Automatically Generated Music

Andrew McLeod[1]

Abstract: In recent years, there has been an explosion of research into the automatic generation of music, both audio and symbolic. Countless deep learning approaches in particular have been proposed, using a wide range of methods and producing an equally wide range of outputs. However, the evaluation of such generations is very difficult, as the gold standard method of evaluation (listening experiments with musically-trained test participants) is expensive, in terms of both time and money (assuming the participants are fairly compensated), particularly when an extensive comparative evaluation is desired. Recent work [Yi23] has undertaken such a procedure, releasing human expert ratings and generated examples comparing human compositions to automatic compositions by several methods. We take the same generations (MIDI files of classical string quartets and piano improvisations), and analyze them instead statistically, comparing properties such as rhythmic density and pitch range across each of the methods and styles. We make no claim that our analysis represents an evaluation of the selected methods, but present our findings as an exploratory look at musically-relevant statistical properties of the outputs of each method, and draw conclusions based on that.

Keywords: Music Generation, Analysis

1 Introduction

The automatic generation of music has been subject to research for many decades. However, in recent years, as deep learning methods became the norm, larger datasets were produced (e.g. MAESTRO [Ha19]), and greater memory became available for training, focus on the task has increased exponentially [BP20]. It should be noted that music generation is a very broad topic, containing within it a wide variety of tasks, including different output formats, and various levels of control and conditioning. In this work, we focus specifically on non-constrained generation of symbolic music, where the style and form of each output is dependent only on the data used to train each model.

Regardless of the type of generation, evaluation including comparison against existing work is not a simple task, due to both the huge number of proposed models and the inherent difficulty of evaluating generated content [YL20]. The best evaluation method is a listening experiment with trained human participants rating the quality of generations, but such an experiment is extremely time consuming and costly (if the participants are compensated).

One recent paper performed such an evaluation, comparing five generation models across two musical styles [Yi23]. The musical styles are Western classical quartets (abbreviated

[1] Fraunhofer IDMT, Ehrenbergstraße 31, 98693 Ilmenau, Germany andrew.mcleod@idmt.fraunhofer.de

as CSQ here and in [Yi23]), lacking microtiming and tempo variation; and classical piano improvisations (abbreviated CPI here and in [Yi23]), performances taken from the MAESTRO dataset [Ha19] which represent live performances of piano compositions, containing expressive timing. We refer the reader to the original publication for further details about the two sets of data. The authors first generated 25 excerpts of 20–30s in each style by each relevant model (some were unable to generate in one of the styles). They then conducted listening experiments with trained human participants. The stimuli were all released publicly for future research.

Here, we take those stimuli and perform a different type of analysis. Rather than trying to estimate the quality of each generation, we measure musically-significant statistical properties (diatonicity, rhythmic texture, etc.). It is important to be clear that this is not a qualitative evaluation, nor an evaluation of any kind. We simply present statistics and discuss similarities and differences between the models rather than remarking on model performance or generation quality directly. Due to lack of space, we do not describe the models (MaMa [CL17], CoRe [Th18], MVAE [Ro18], MuTr [Hu18], LiTr[2]) here, and rather refer readers to the original paper for an overview.

2 Analysis

This section presents our analyses for each of the models in each style. CSQ includes two sets of 25 generations from human composers: one set from Haydn, Mozart, and Beethoven (Orig) matching the training style, and one set from Vivaldi and Brahms (BeAf) from before and after the era of the training data. CPI includes a set of 25 human compositions (Orig) matching the training style. We divide our analyses into two categories: rhythmic, which concern features related to note onset and offset positions; and pitch-based, which involve the pitch of each generated note. With the exception of LiTr (which was used pre-trained on a large corpus of scraped data), all models were trained from scratch on data in each style.

We visualize the analyses in violin plots. The white dots represent the median, while the thicker black bars range from the first to the third quartile of the data. The thickness of the colorful shape represents the distribution of values, smoothed using kernel density estimation. For each analysis, a generation is a single point in the distribution (e.g., for average duration, each point is the average duration of all notes in a single generation).

2.1 Rhythmic

First, we measure the average polyphony level of each generation (i.e., the average number of simultaneous notes at each point in time in each generation). The results are plotted in Figure 1, where it is clear that, for CSQ, MaMa, MVAE, and MuTr model the original data

[2] https://magenta.github.io/listen-to-transformer

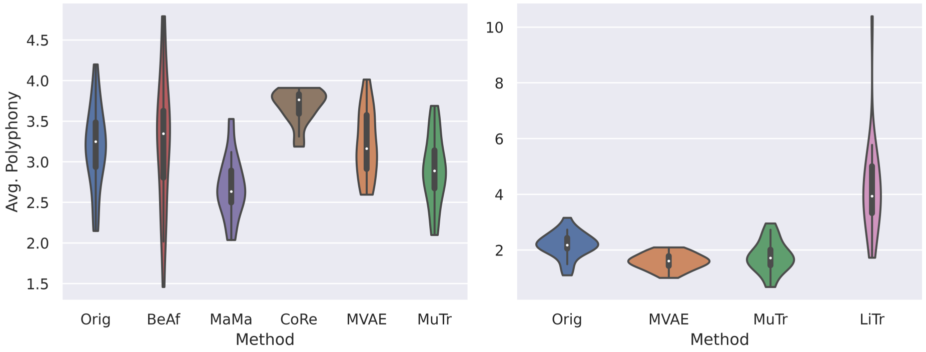

Fig. 1: The average polyphony level of each model on CSQ (left) and CPI (right).

most closely in this regard, with CoRe generating more music towards the high end of the training data's polyphonic. CoRe's explicit modeling of each voice may be the cause of this, as it is less likely to generate silences within each voice than a model without explicit voices. For CPI, both MVAE and MuTr modeled the training data's polyphony quite well, with MuTr again slightly closer to the original distribution. LiTr is a clear outlier, generating pieces with much greater polyphony due to its training data.

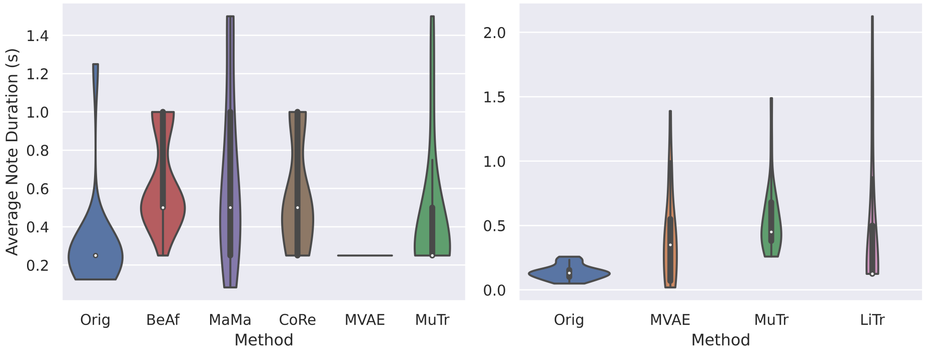

Fig. 2: The note duration of each model on CSQ (left) and CPI (right).

We next present the average note duration for each method in Figure 2. For CSQ, a quarter note is 0.5 seconds, and it can be seen that by far the most common average duration in the original style is around an eighth note (and MVAE produced only music with exactly that average duration), although there is a long tail towards higher durations. MaMa produced the widest range of average durations, roughly evenly distributed across the spectrum from 16th notes to dotted half notes, while none of the other methods generated a single piece whose average duration was less than a quarter note. For CPI, the original dataset contains much shorter average durations (with a maximum of around 0.25 seconds) than all of the models, which each produced a wide range of average durations, up to over 1 second.

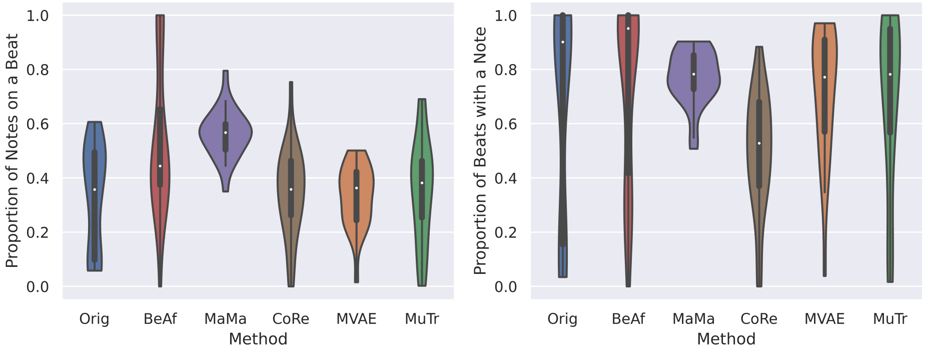

Fig. 3: The average proportion of notes that lie on a beat (left) and the average number of beats which have at least one note onset (right) for each model for the CSQ data.

Finally, in Figure 3, we present two features which are only informative for outputs which don't include expressive timing (in our case, CSQ), and generally describe the metrical density and regularity of the generation. In these plots, MaMa is the clear outlier, generating a much greater proportion of its notes on beat, and tending to produce only generations where around 80% of beats have at least one note. It's possible that, similar to CoRe with polyphony, the explicit modeling of beat position in MaMa's state space induces this property, whereas the other models all exhibit distributions closer to the original style.

2.2 Pitch-based

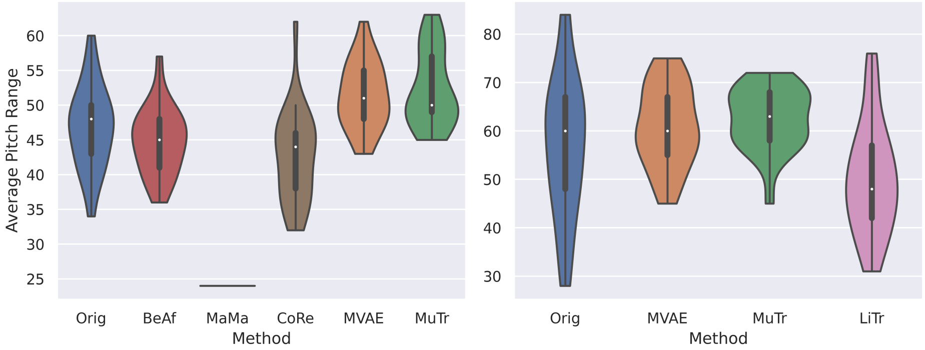

Fig. 4: The average pitch range of each model on CSQ (left) and CPI (right).

We first measure the average pitch range of each method in Figure 4. For CSQ, we see that MaMa is again an outlier, generating only music with a pitch range of 25. Again, this is likely due to a limitation of the model itself. For the deep learning methods, MVAE and MuTr both tend to generate music with slightly wider pitch ranges than the original data, while CoRe (which explicitly models the relationship between the notes in each voice), matches the training distribution more closely. It seems that CoRe's modeling strategy has helped it to implicitly capture a pitch range dependency that the less constrained models missed. For CPI, the generations all lie within the range of the original data, although MVAE and MuTr are again more towards the upper-end of the distribution. In the case of pitch range, unconstrained models do not yet capture the dependence between high and low pitches that CoRe explicitly models.

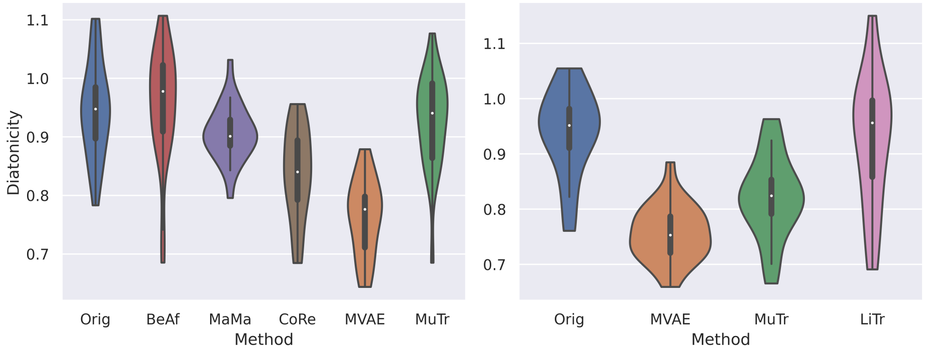

Fig. 5: The estimated diatonicity (based on overlap with picth-class-profile templates from [TM08]) of each model on CSQ (left) and CPI (right).

Finally, we measure the level of diatonicity of each method, shown in Figure 5, by first measuring each piece's normalized pitch class profile (PCP). That PCP is then multiplied by the major and minor template PCPs from [TM08], rotated from 0 to 11 times to model different tonics. The maximum of these 24 products is taken as a piece's diatonicity. A PCP that exactly matches one of the rotations of a template would have a value of 1, while larger values represent pieces that tend to use more of the most common pitches of a template. Here, MVAE (and to a lesser extend CoRe) is a clear outlier for both CSQ and CPI, generating music that is less diatonic than the original style. The ability to stay within a tonal center relies on long-term modeling of pitch, something that the RNNs of CoRe and MVAE tend to struggle with compared to Transformers. MaMa models pitches explicitly relative to an estimated tonic, and thus needs no long-term model to remain in the same key.

3 Conclusion

In this paper, we have presented a musically-informed statistical analysis of some of the state-of-the-art systems for music generation. We used the generations produced and released by [Yi23], covering both quantized outputs (i.e., compositions) in the style of Classical string quartets, and expressive performances generated in the style of classical piano improvisations. We measured properties of the resulting generations, producing both rhythmic and pitch-based analyses.

In general, we found that the generated examples tend to match the original data in the measured features. However, there were some interesting exceptions. Average duration was the most different for all methods compared to the original distribution, showing that perhaps rhythmic properties of the music are more difficult for the current models to learn. We also showed that for MVAE and CoRe (both forms of RNN for which long-term dependence is difficult to model), remaining in the same key for the entire generation was uncommon, which points to a potential drawback of such architectures. Finally, in two cases where a method explicitly models one aspect of the music (CoRe for polyphony and MaMa for metric regularity), they produced outputs that were skewed towards one end of the training distribution, suggesting that such explicit modeling, though helpful in some cases (e.g., MaMa's diatonicity), can also lead to an over-regularity of that feature in the generations.

Bibliography

[BP20] Briot, Jean-Pierre; Pachet, François: Deep learning for music generation: challenges and directions. Neural Computing and Applications, 32(4):981–993, 2020.

[CL17] Collins, Tom; Laney, Robin: Computer-generated stylistic compositions with long-term repetitive and phrasal structure. Journal of Creative Music Systems, 1(2), 2017.

[Ha19] Hawthorne, Curtis; Stasyuk, Andriy; Roberts, Adam; Simon, Ian; Huang, Cheng-Zhi Anna; Dieleman, Sander; Elsen, Erich; Engel, Jesse; Eck, Douglas: Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In: ICLR. 2019.

[Hu18] Huang, Cheng-Zhi Anna; Vaswani, Ashish; Uszkoreit, Jakob; Shazeer, Noam; Simon, Ian; Hawthorne, Curtis; Dai, Andrew M.; Hoffman, Matthew D.; Dinculescu, Monica; Eck, Douglas: Music Transformer. In: ICLR. 2018.

[Ro18] Roberts, Adam; Engel, Jesse; Raffel, Colin; Hawthorne, Curtis; Eck, Douglas: A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In: ICML. pp. 6939–6954, 2018.

[Th18] Thickstun, John; Harchaoui, Zaid; Foster, Dean P; Kakade, Sham M: Coupled recurrent models for polyphonic music composition. In: ISMIR. 2018.

[TM08] Temperley, David; Marvin, Elizabeth West: Pitch-class distribution and the identification of key. Music Perception, 25(3):193–212, 2008.

[Yi23] Yin, Zongyu; Reuben, Federico; Stepney, Susan; Collins, Tom: Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. Machine Learning, pp. 1–38, 3 2023.

[YL20] Yang, Li-Chia; Lerch, Alexander: On the evaluation of generative models in music. Neural Computing and Applications, 32(9):4773–4784, 2020.