Augmenting Conditional Music Generation

Brandon Tiu
Department of Computer Science
University of Toronto
27 King's College Cir, Toronto, ON M5S 1A1
brandon.tiu@mail.utoronto.ca

David Basil
Department of Computer Science
University of Toronto
27 King's College Cir, Toronto, ON M5S 1A1
david.basil@mail.utoronto.ca

Isaac Wetmore
Department of Computer Science
University of Toronto
27 King's College Cir, Toronto, ON M5S 1A1
isaac.wetmore@mail.utoronto.ca

Shysta Sehgal
Department of Computer Science
University of Toronto
27 King's College Cir, Toronto, ON M5S 1A1
shysta.sehgal@mail.utoronto.ca

Abstract

Recent advances in deep music generation frequently involve training on extensive datasets comprising thousands of hours of music. However, despite the availability of such data, many contemporary models, such as MetaAI's MusicGen, often forego the implementation of data augmentation techniques during pre-processing. This oversight raises questions about the robustness and flexibility of the models when confronted with modified input data. This study aims to bridge this gap by conducting experiments that employ diverse augmentation techniques, including pitch shifting, time stretching, and volume modification, to critically assess the model's ability to maintain musical coherence and quality. By presenting both original and systematically augmented datasets to MusicGen, we evaluate the model's generative performance through quantitative metrics such as Root Mean Squared Error and Mel-frequency Cepstral Coefficients, as well as qualitative human judgment via a Turing-like test. The findings suggest that while the model demonstrates a basic competency in generating continuations from unaltered prompts, its ability to adapt to and correct for augmented inputs is limited. These results underscore the necessity for incorporating resilience to data imperfections as a key criterion in the evolution of deep learning models for music generation.

1 Introduction

Self-attention Transformers [Vaswani, et al., 2017] have taken the world by storm with their ability to take in data from numerous differently input modalities, such as text, images, or video, and accurately generate sequential data that follows those given prompts. One area of research with ample potential is the ability for these models to generate music. Just as an LLM takes in text and can generate whatever sequence is most likely to follow, can we input a musical piece into a model built off a self-attention mechanism, and generate a reasonable continuation of that music prompt?

Research into this question has been fast developing, with notable attempts by researchers at places like Facebook (Copet, et al., 2023) and Google (Huang, et al., 2018) providing potential models that generate music based off provided prompt music. These proposed models have shown preliminary success in being able to generate music that follows an expected structure, and flows nicely from the given prompt; however, exploration into the quality of these models remains young. This paper seeks to provide some much-needed strain on these models, testing how notable augmentations to original music prompts can impact the effectiveness of these models.

We take MusicGen, a model created by (Copet, et al., 2023), and provide an extension of their work by making two extensions to our prompts. The first is the addition of more music that involves more than just piano music, giving the model more instruments to decode and generate. We also apply augmentations related to factors such as tempo, pitch, and volume, and measure the success of the models' ability to generate coherent music as a follow up to a prompt. To evaluate the performance of a model, we use three metrics: a RMSE comparison of the output song with the prompt provided; a metric described known as MFCC, which measures the impact of music distortions on original music, and human evaluation, employing a Turing test to gauge whether participants can note the difference between the full original piece, and the music created by the MusicGen model.

2 Problem Formulation and Motivation

A big strength of these text and image-based models that use Transformer architecture—such as ChatGPT—comes from how adaptive and resilient these models are. For example, imperfect text prompts, such as prompts filled with grammar mistakes or typos, can be readily understood by large language models such as ChatGPT without much trouble. Images with slightly blurry pixels can still be readily classified by models such as DALLE (Ramesh, et al., 2021). Finding success with perfectly constructed prompts is one thing, but a model's performance should also be gauged by its ability to adapt to imperfect inputs with the same amount of ease one would expect a human to easily sidestep minor errors of a query.

Our research seeks to simulate this problem of imperfect prompts but in the context of music generation. We do so by proposing the following research problem: to what extent can these models, when presented with arguments or distorted music prompts, still generate music that is both consistent with the prompt (a clear continuation of the music given by the prompt) but also correcting, that is, a generated output that maintains structure found in more traditional music pieces. If these models are not resilient, minor distortions of input may generate music with structure notably divergent to the input, suggesting the models poorly generalize when given imperfect data. If an argumentation as simple as changing the volume of a piece seriously corrupts the output, then these inflexible models have a long way to go before they can be used in everyday life, where imperfect data is a constant fact of life.

3 Prior Work

Music generation based on self-attention began with the work of (Huang, et al., 2018). Their research explored the capabilities of these models not only in generating music but also in assessing their responsiveness to minor alterations in the input data. Their study served as a limited introduction into the topic of music augmentation however, it was confined solely to two augmentations–pitch and time interval. Furthermore their inquiry was restricted to single instrument music specifically solo piano compositions.

Another paper by (Godwin, et al., 2021) also explored the impact data augmentation had on the results of generative models, but took a much more subjective approach. Instead of augmenting by objective measures, such as pitch, they made distortions in music with the intention of shifting the mood or tone a piece conveyed. By changing the emotional weight of a piece, they observed whether the model could respond to such changes, noting whether the music generated in response also changed its emotional tone, or was unable to generalize to different emotional moods. While this approach is interesting, it's ultimately dependent on a lot of subjective measures for shifting the emotional weight of a piece. Furthermore, these types of augmentations underline a fundamentally different objective. While augmentations like this explore how music generation models can generalize to music of different genres and tones, our paper asks whether these music generation models are resilient against deliberate distortions and unpredictable shifts in music characteristics.

Even in the last four years, the progress these models have made is impressive. But as these models become more advanced, we note a decrease in work that incorporates augmentation into its analysis. For Meta's model, (Copet, et al., 2023) no augmentation is applied for music prompts, instead choosing to focus on only augmenting text prompts. Our paper seeks to contribute to the research behind music augmentation, proposing a methodology for testing music generation algorithms with objective augmentation metrics that can thus be analyzed with objective analysis metrics.

4 Methods and Experimental Setup

4.1 Dataset

Music data for prompting was sampled from the Lakh MIDI dataset (Raffel, 2016), which contains music from numerous genres and instruments. The complexity of this music goes well beyond the basic piano music found in many studies, providing an extra level of strain not common in evaluation research. Out of the over 40,000 songs extracted from the data, we sampled 100 MIDI files to serve as prompts to our model.

4.2 Data Augmentations

In our study, we focused on the application of data augmentation techniques to MIDI files, selecting three primary augmentations: pitch, volume, and tempo. Augmentations were applied by sampling from uniform distributions encompassing varying ranges. Specifically, pitch alterations were sampled from a uniform distribution within the interval [-6, 6] in semitones, while volume adjustments were drawn from a uniform distribution spanning [-30, 30] on the midi scale between 0-127. Tempo modifications were executed by sampling from a discrete distribution comprising values of [0.25, 0.5, 0.75, 1.25, 1.5, 1.75], where, for instance, a value of 0.25 signifies a reduction in tempo by a factor of 0.25 for each note.

4.3 Model

Our investigation expands upon the framework presented by (Copet, et al., 2023), denoted as MusicGen. We specifically harness the capabilities of the melody checkpoint within the 1.5B model variant, facilitating dual conditioning with text and melody inputs during the generative process. While the primary objective of MusicGen pertained to the examination of music and text prompt interactions, our study confines its scope to the decoding proficiency with music prompts exclusively. The model leverages recent advancements in the field of music generation models, thereby establishing itself as a suitable candidate for evaluating our proposed data augmentation techniques. Employing a self-attending transformer architecture, it embodies contemporary methodologies in the domain.

4.4 Model Conditioning

Before feeding in our augmented MIDI files, we first synthesized our results (using FluidSynth) into a .wav file. We then applied trimmed the extracted melody of the file to a length of 5 seconds. This means that as an input, the model received 5 seconds of a prompt. It then generated what it expected to be the next 5 seconds of the music, leading to the creation of a 10 second file for analysis. This .wav output was then applied to three different metrics described below.

4.5 Root Mean Squared Error

To quantify the deviation of generated outputs from the standard output with no augmentations applied, we compute the root mean squared error (RMSE) between the generated outputs. We used the torchaudio library to load the generated audio clips as waveforms. RMSE was selected as the evaluation metric due to its ability to provide an error value in the same units as the audio samples (e.g., decibels), facilitating a direct comprehension of the magnitude of differences between the samples. Furthermore, RMSE inherently penalizes large errors more severely due to the square root operation. This characteristic is valuable when comparing music samples because significant discrepancies in amplitude or phase can lead to perceptually noticeable differences in the sound, which should be reflected in the evaluation metric.

Prior to RMSE calculation, we ensure that the generated waveforms possess identical sample rates and lengths. Additionally, all waveforms undergo normalization to standardize their scales, thereby mitigating any amplitude differentials among them. Such standardization is essential to prevent discrepancies in overall amplitude from disproportionately affecting similarity measurements, which could obscure similarities in waveform shapes.

4.6 MFCC Evaluation

Quantitative evaluation of written music is quite difficult, naturally. We aimed to cover a number of bases in our evaluation methods. Our evaluation using MFCC was intended to be a measure of how 'faithful' the continuations were to the musical prompts.

MFCC, or Mel-frequency cepstral coefficients, are representations of the power spectrum of an audio clip. They are widely used on speech recognition task, as it is largely believed that they capture some of the most important aspects of sound for human comprehension, especially timbre, which can be an important aspect of musical style. Thus, they seemed well fit. We took each generated piece of music as well as the piece that it was meant to be a continuation of and found their MFCCs. They were then compared to see how well they each acted as a continuation of the prompt.

We employed the torchaudio library's method to compute the MFCC. The inputs were converted to mono to match the outputs. Each input audio clip was five seconds and most continuations were around that timeframe too. Any that weren't, were trimmed to be the same length as the input it was being compared against. The question of length became an especially important one with regards to the tempo augmentation. When a five second clip is augmented to have half the tempo, it has now become a ten second clip. Since we wanted the model to have a similar number of notes (and thus, a similar amount of information) with each augmentation, we did not force the tempo-augmented examples to be five seconds each. This means that some samples are longer than others, and will affect the evaluation of the similarity between the two resultant MFCC vectors. It prompted us to examine cosine similarity as well as our original intention of euclidean distance, as the former is theoretically more applicable when comparing the differences between vectors of different dimensionalities.

4.7 Musical Turing Test Evaluation

To evaluate the model using human feedback, we asked 10 participants to answer the questions in Table 1. The questions were inspired by Hernandez-Olivan, et al. (2022).

Table 1: Survey questions

Question	Score Range
I feel I have heard similar music before	1-5
The piece conveys something to me	1-5
Is this sample composed by a human or is it AI-generated?	AI/H/NS
Musical composition overall rating	1-5

Participants were divided into two groups, each group exposed to a distinct set of music samples. Each set consisted of ten samples, with six composed by AI and four by human composers. These samples also contained the augmentations of the original sample as well as music generated by the model on the augmented samples.

Participants were provided with a Google Form where they rated each music sample on familiarity, emotional impact, and overall quality on a scale of 1 to 5. They also indicated whether they believed the composition was created by AI, a human, or if they were not sure.

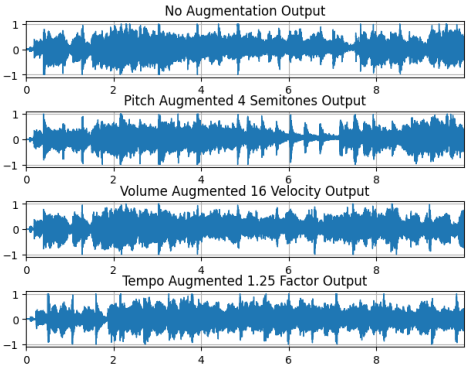
4.7.1 Data Analysis

We replaced the 'Not sure' responses with NaN for appropriate handling in statistical tests. Descriptive statistics were calculated for Questions 1, 2, and 4, both within and across sets. The Shapiro-Wilk test assessed normality, and the Mann-Whitney U test compared the AI and human scores, both within and across sets. The Spearman correlation analyzed relationships between the questions.

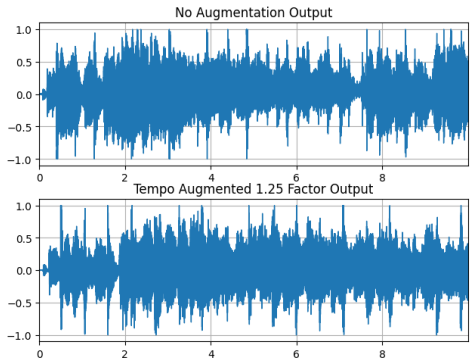
5 Results

5.1 Comparison with the standard output

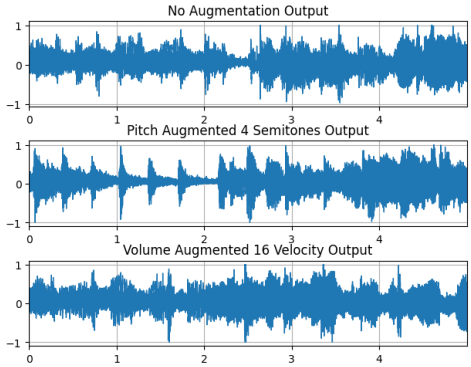
Figure 1 presents the visualization of the waveforms of the outputs for one audio prompt. We can see the effects of the augmentation reflected on the first 5 seconds in Figure 1a but that the prompt



(a) All outputs



(b) Standard vs tempo augmentation



(c) Standard vs pitch vs volume augmentation

Figure 1: Comparative analysis of MusicGen outputs conditioned using a 5 second prompt under varying data augmentation strategies.

Table 2: The root mean squared error between an output conditioned on the given audio prompt with no augmentation and selected augmentation averaged across all generated examples.

Augmentation	Average Root Mean Squared Error
Pitch	0.313
Volume	0.310
Tempo	0.293

remains largely unaltered. Conversely, the generated 5 seconds exhibits significant variation, as depicted in Figure 1c. Regarding tempo augmentations we observe that the generated output closely resembles the standard output.

As the initial 5 second prompts often demonstrate high similarity for pitch and volume augmentations, we specifically computed the RMSE on the successive generated segments. Since tempo augmentations essentially change the length of the prompt we calculated the RMSE on the full 10 seconds. After averaging across all the examined prompts with randomly applied augmentations we observe in Table 2 that pitch and volume augmentations tend to contribute a larger deviation from what would be the standard output without augmentations applied vs tempo augmentations.

5.2 MFCC Analysis

Evaluating the similarities of MFCCs using Euclidean distance rounded to a whole number, we get the results in Table 3

Using cosine similarity, we get the results in Table 4

Table 3: Similarities of MFCC using Euclidean distance

Augmentation	None	Pitch	Tempo	Volume
Min L2	2877	2659	1978	2161
Max L2	12957	9662	11292	10335
Mean L2	4334	4405	4169	4447

Table 4: Similarities of MFCC using cosine similarity

Augmentation	None	Pitch	Tempo	Volume
Min CS	-0.55	-0.36	-0.31	-0.31
Max CS	0.82	0.99	0.99	0.99
Mean CS	0.39	0.34	0.32	0.29

The MFCC results suggest that augmentation makes the quality of the model's continuations worse. We see that, on average, the most 'faithful' continuations, as measured by MFCC similarity, are those where the data has not undergone augmentation. This is true when measured with cosine similarity. Using L2 distance to infer similarity, it seems the tempo-augmented data performs best. However this may likely be chalked up to variability in the samples' lengths caused by speeding up or slowing down the samples. This is supported by the maximum and minimum L2 values for tempo augmented data being outside the range of data for other kinds of augmentation, suggesting wide variability.

On the other hand, we see that augmented data has higher maximum cosine similarity than non-augmented data. It would be worthwhile to attempt to understand why this is: whether the model is unable to 'continue' the piece in a meaningful way, so it just repeats what it has heard so far, or whether some augmentations are useful for the way the model encodes information. Data with no augmentation has the lowest minimum cosine singularity, which may give support to the theory that the model is willing to make more 'daring choices' with regards to unaugmented data, while it tries to just repeat augmented data back out.

5.3 Musical Turing Test Results

5.3.1 Descriptive Statistics

Set 1 exhibited a mean familiarity score (Question 1) of 3.38 (SD = 1.34), an emotional impact score (Question 2) of 2.60 (SD = 1.67), and an overall rating (Question 4) of 2.56 (SD = 1.25). Set 2 presented lower mean scores with a mean familiarity score of 2.80 (SD = 1.21), emotional impact score of 2.44 (SD = 1.03), and an overall rating of 2.40 (SD = 1.11). Across both sets, the mean scores were 3.09 (SD = 1.30) for familiarity, 2.52 (SD = 1.38) for emotional impact, and 2.48 (SD = 1.18) for overall rating.

Table 5: Descriptive statistics for Set 1.

Statistic	Question 1	Question 2	Question 4
Count	50	50	50
Mean	3.38	2.60	2.56
Std. Deviation	1.34	1.67	1.25
Minimum	1.00	1.00	1.00
25th Percentile	2.25	1.00	1.00
Median	3.00	2.00	3.00
75th Percentile	5.00	4.00	3.00
Maximum	5.00	5.00	5.00

5.3.2 Inferential Statistics

Shapiro-Wilk tests for normality indicated that the ratings for both AI and human compositions did not follow a normal distribution, with $p < .05$ for both groups, which is why we chose Mann-Whitney U test for our purposes.

Table 6: Descriptive statistics for Set 2.

Statistic	Question 1	Question 2	Question 4
Count	50	50	50
Mean	2.80	2.44	2.40
Std. Deviation	1.21	1.03	1.11
Minimum	1.00	1.00	1.00
25th Percentile	2.00	2.00	2.00
Median	3.00	2.50	2.00
75th Percentile	3.75	3.00	3.00
Maximum	5.00	5.00	5.00

Table 7: Descriptive statistics for all sets combined.

Statistic	Question 1	Question 2	Question 4
Count	100	100	100
Mean	3.09	2.52	2.48
Std. Deviation	1.30	1.38	1.18
Minimum	1.00	1.00	1.00
25th Percentile	2.00	1.00	1.00
Median	3.00	2.00	2.50
75th Percentile	4.00	4.00	3.00
Maximum	5.00	5.00	5.00

The Mann-Whitney U test was conducted to evaluate differences in the rating distributions between AI-composed and human-composed music within each set and across the entire dataset. For Set 1, the test revealed a statistically significant difference ($U = 27.5, p < .001$), suggesting that participants could distinguish between AI and human compositions, showing a preference or distinct perception for one type over the other. In contrast, within Set 2, the test showed no significant difference in the rating distributions ($U = 172.5, p = .163$), indicating a more ambiguous distinction between the compositions, possibly due to the nature of the pieces or the participants' interpretations.

The median scores suggested potential interaction effects, with the median rating for AI compositions increasing from 1.5 in Set 1 to 2.0 in Set 2, and the median rating for Human compositions decreasing from 4.0 to 3.0 across the sets.

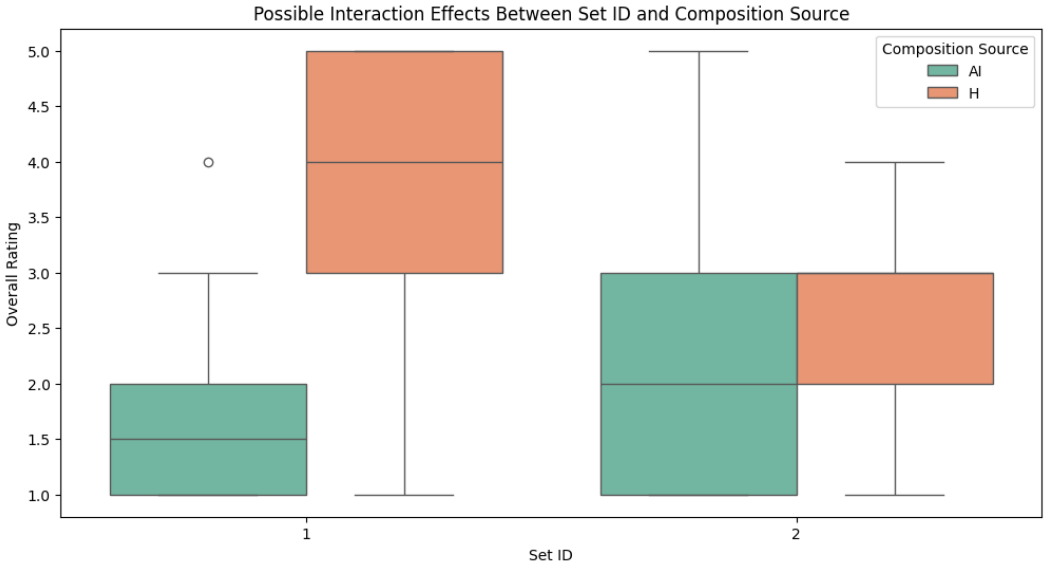


Figure 2: Possible interaction effects between set ID and composition source

When the data from both sets were combined, the significant difference reemerged ($U = 350.5, p < .001$), highlighting a consistent pattern across the entire sample despite the anomaly in Set 2. This overall difference across sets suggests a general ability to differentiate between the sources of the compositions when considering a larger sample size.

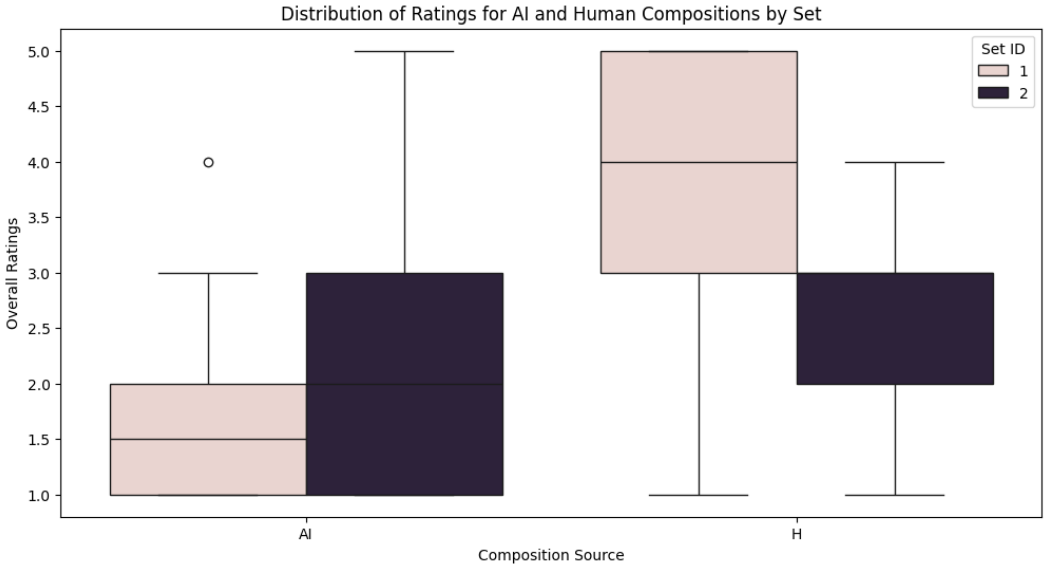


Figure 3: Distribution of ratings for AI and human compositions by set

5.3.3 Correlation Analysis

Spearman correlation coefficients showed a moderate positive relationship between familiarity (Question 1) and emotional impact (Question 2) ($r_s = .501$), and between emotional impact (Question 2) and overall rating (Question 4) ($r_s = .528$). A weaker positive relationship was observed between familiarity and overall rating ($r_s = .283$).

6 Discussion and Conclusion

Results from our MFCC and MSE analysis suggest when pitch and volume are applied as augmentation, we generally observe lower performance as compared to the original music prompt. Tempo produces more erratic results, but this may be the result of limitations with our methodology (explained below in limitations). More insight comes from our Turing test, in which participants reported much weaker quality scores for music generated by MusicGen as compared to human produces pieces.

These results suggest preliminary evidence that augmentation has a distinctly negative impact on the quality of MusicGen in producing music. While previous studies have explored augmentations related to mood or style, these augmentations tend to be variations of music that are still within an predicted or standard expected input. Varying a piece by making its mood more sad, for example, likely wouldn't largely impact a model's performance, as it could then just draw upon its exposure to sad music as a reference. Our study differed by exploring more abnormal augmentations that the model likely was not exposed to during training. The Turing test results are especially telling, as participants still rated the human generated, but augmented music relatively high, while rating the generated music significantly lower. This suggests the model deviates quite significantly from the original piece when even minor distortions are applied to the music. This divergence, of the model MusicGen and potentially other music generation models, suggests these models may struggle with generalizability.

Limitations to this study involve tempo augmentations, and urge further exploration into how it influences music generation. Part of the problem is likely our inability to scale the length of the prompt music by the tempo to provide the model with an equal amount of music to generate from.

The biggest limitation here was computing resources–providing the model with more music would have meant also requesting the model to generate more music, requiring even more computation for a computationally extensive task. Another limitation was the small sample size of our Turing Test, suggesting our results should be taken as preliminary evidence. More research should be done on not only how we we should approach music augmentation, but also how we should be evaluating music generation models in general. Our paper attempts to provide a framework for a more objective analysis of music generation.

7 Contributions

Brandon contributed to loading the model checkpoint, generating outputs, and doing the RMSE analysis. Isaac contributed to parsing and formatting the Lakh dataset, generating and formatting model outputs, augmenting tempo, and writing first 3 sections and part of section 4 and the discussion. David contributed to augmenting by volume and evaluating with MFCC. Shysta contributed to augmenting by pitch and conducting the Turing test evaluations and performing analysis on the collected data. Everyone contributed to the report.

References

[1] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2024) Simple and Controllable Music Generation. arXiv:2306.05284 [cs.SD].

[2] Godwin, T., Rizos, G., Baird, A., Al Futaisi, N. D., Brisse, V., & Schuller, B. W. (2021) Evaluating Deep Music Generation Methods Using Data Augmentation. arXiv:2201.00052 [cs.SD].

[3] Hernandez-Olivan, C., Abadias Puyuelo, J., & Beltran, J. R. (2022). Subjective Evaluation of Deep Learning Models for Symbolic Music Composition. arXiv:2203.14641v2 [cs.SD].

[4] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018) Music Transformer. arXiv:1809.04281 [cs.LG].

[5] Peter Hanappe, Conrad Berhörster, Antoine Schmitt, Pedro López-Cabanillas, Josh Green, David Henningsson, and Tom Moebert. FluidSynth: A Real-Time Software Synthesizer. Version 2.3.5, January 11, 2024. [Online]. Available: https://www.fluidsynth.org

[6] Raffel, Colin (2016) Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching.

[7] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021) Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV].

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023) Attention Is All You Need. arXiv:1706.03762 [cs.CL].