Comp 138: Reinforcement Learning
Tufts University, Fall 2023

Emmet Hayes
Final Project

Reinforcement Learning for Musical Composition: A Tabular Approach

Section 1: Introduction

1.1 Inspiration and Motivation

The invention of the player piano in 1896 marked a revolutionary step in the history of musical expression. These intricate machines could mechanically reproduce a piece of music by following perforations on a roll of paper known as a piano roll. As the roll passed over a tracker bar, it activated pneumatic devices that, in turn, set wooden fingers in motion to strike the corresponding notes on the keyboard. This mechanism allowed for the replication of complex performances, even transcending the limitations of human players.

Significantly, player pianos introduced the concept of encoding music in a form that was independent of the performer. This encoding was not just a replication of notes but also encapsulated nuances like tempo, loudness, and other dynamics, controlled by levers and pedals. It represented an early form of storing and reproducing human creativity and performance, a precursor to modern digital formats.

In the contemporary era, this concept has been transformed and expanded through the advent of MIDI (Musical Instrument Digital Interface). MIDI is essentially the digital descendent of the piano roll, encapsulating the essence of a performance in a format abstracted from the physical instrument. This technology has democratized music production, allowing artists to compose, edit, and share their work with unprecedented ease and flexibility. MIDI files have become ubiquitous across the internet. They span an impressive array of genres and styles, and are often available free of charge. This accessibility has fostered a diverse community of music composers, enthusiasts, and learners, significantly altering the landscape of music creation and distribution.

The application of Reinforcement Learning (RL) techniques to generate these digital rolls presents an exciting frontier in music technology. While there is substantial research in other domains of RL, its application in music, especially in MIDI generation, is relatively nascent. Approaches like LSTM-based DQNs (Long Short Term Memory-based Deep-Q-Networks), explored in various initiatives, such as Google's Magenta project, demonstrate the potential of machine learning in capturing and creating complex musical patterns. Incorporating human feedback into the reward systems of these RL techniques opens up possibilities for creating custom interpolations, bridging the gap between human creativity and algorithmic efficiency. This blend of human artistry and machine precision could lead to novel forms of musical expression. Thus, by exploring the exciting mechanisms inspired from mechanical piano rolls of the past, creative tools can be devised to augment human creativity, offering new avenues for artistic expression and experimentation.

1.2 Why MIDI and Not Raw Audio?

The choice of Musical Instrument Digital Interface (MIDI) over raw audio as the primary data format for reinforcement learning in music composition warrants a detailed examination. MIDI, an established standard in musical notation, offers a structured and rich representation of musical compositions that surpasses the capabilities of raw audio in several critical aspects.

- 1. ***Structured Representation of Musical Elements***: MIDI provides a precise and well-defined format for encoding musical elements. This includes the frequency of notes, their duration, articulation, and a plethora of other vital metadata. Such a structured representation is essential for a reinforcement learning agent tasked with learning and mimicking human compositions. In contrast, raw audio lacks this direct representation of musical elements, often rendering it an inefficient choice for learning detailed compositional structures.

- 2. ***Discrete Encoding of Musical Information***: MIDI excels in discretely representing critical aspects of music. For instance, pitch values in MIDI are stored in a 7-bit format, offering a range from 0 to 127. This granularity allows for a nuanced understanding and generation of pitch variations, a key component in music composition. Moreover, MIDI files encapsulate other significant performance details like note duration, time steps, voice settings, and velocity (dynamics). Such discrete storage of data is instrumental for algorithms to parse and learn from complex musical compositions.

- 3. ***Individual Instrument Encoding***: One of the standout features of MIDI is its capability to encode details of each instrument's performance separately in a multi-instrumental composition. This is in stark contrast to raw audio, where the sound is typically a sum combination of all instruments, making individual instrument analysis more challenging. This feature of MIDI is particularly beneficial for reinforcement learning models aimed at understanding and replicating the complexities of ensemble performances.

- 4. ***Conversion Efficiency to Raw Audio***: Converting MIDI to raw audio is a straightforward process, achievable with various Digital Audio Workstation software, such as Garageband, Logic, Ableton, Pro Tools, FL Studio, and Audacity. This ease of conversion facilitates the practical application of learned compositions in a format that is universally accessible and recognizable.

- 5. ***Challenges in Reverse Conversion***: In contrast, generating MIDI from raw audio presents significant challenges. An audio file, such as an MP3 or WAV, typically contains a 2D float array representing a composite audio signal from multiple instruments. While there are AI solutions attempting this conversion (including features in software like Ableton), they often fall short in accurately dissecting and representing multi-instrumental compositions in MIDI format. This limitation further underscores the practicality of using MIDI as the starting point for reinforcement learning in music.

In summary, the selection of MIDI over raw audio is rooted in its structured, discrete, and detailed representation of musical elements, which are pivotal for training reinforcement learning agents in music composition. This choice not only aligns with the technical requirements of such learning models but also significantly enhances the quality and accuracy of the generated compositions.

1.3 Convergence Conditions of Off-Policy Monte Carlo Control in Music Generation

Off-policy Monte Carlo control using importance sampling is a powerful method in reinforcement learning, especially suited for complex tasks like music generation. It is discussed in detail in this famous textbook, written by Sutton and Barto [1]. This approach allows the learning agent to evaluate and improve a policy different from the one it follows, providing flexibility and robustness in learning. The convergence of this method in the context of music generation relies on several critical conditions:

- 1. ***Importance Sampling***: The core of off-policy Monte Carlo control is importance sampling, a technique that adjusts the expectation based on the probability discrepancy between the target and behavior policies. For convergence in music generation, it's vital that the importance sampling ratios are properly calculated and applied. This ensures that the estimates are unbiased and converge to the true value of the target policy, even though the data is generated under a different behavior policy.

- 2. ***Consistent Exploration Policy***: The behavior policy must ensure continual exploration of the action space. In the context of music composition, this means the policy must occasionally select actions that deviate from the current best-known strategy for composing music. This could involve exploring unconventional chord progressions, rhythms, or instrumentations. Adequate exploration is crucial for the agent to learn the full range of possible musical compositions and avoid premature convergence to suboptimal policies.

- 3. ***Stable and Adequate Coverage***: The target policy, which is being improved, must be covered by the behavior policy. That is, every action that has a non-zero probability under the target policy must also have a non-zero probability under the behavior policy. In musical terms, the behavior policy must be capable of exploring any sequence of notes or rhythms that the target policy might consider. This condition is essential to ensure that the importance sampling ratios remain well-defined and finite.

- 4. ***Convergence of Weighted Importance Sampling Estimates***: The weighted importance sampling estimates must converge. This requires that the variances of the importance sampling ratios are controlled and do not grow too large. In music generation, this might translate to ensuring that the behavior policy does not deviate too drastically from the target policy, as extreme deviations can lead to high variance in the estimates and hinder convergence.

- 5. ***Robust Sampling Strategy***: Given the complexity and variability inherent in musical composition, the sampling strategy must be robust enough to handle the diverse range of musical structures and styles. The agent should be capable of handling various genres, tempos, and instrumentations without losing the ability to learn effectively.

The application of off-policy Monte Carlo control using importance sampling in music generation requires careful attention to the principles of importance sampling, consistent exploration, policy coverage, control of variance in importance sampling ratios, and a robust sampling strategy. Adhering to these conditions is vital for ensuring the effective convergence of the algorithm, enabling it to learn a wide variety of musical compositions effectively and creatively.

Section 2: Problem Approach

2.1 Noteworthy Domain-Specific Utility Functions

The development of domain-specific utility functions played a pivotal role in bridging the gap between the MIDI format and the reinforcement learning model. These functions were essential for converting MIDI data into a format suitable for the model and translating the model's output back into MIDI. Their functionality and significance are outlined as follows:

- 1. ***MIDI to State Sequence Conversion***: One of the key utility functions involved transforming MIDI data into a sequence of states. This process began with the invocation of a pre-trained Magenta model, which generated a simple melody in MIDI format. The MIDI data, typically consisting of note pitches, durations, and start times, was then converted into a list of lists. Each list represented a state in the context of the reduced state space, encapsulating the dimensions of pitch, duration, and start time. This sequence of states effectively formed the goal states in the reinforcement learning environment. When the agent played notes or timings that matched this goal sequence, it was rewarded, reinforcing the learning of musical patterns that align with the generated melody. Alternatively, included are utility functions for generating sequences based on commonly-used scales as defined in Classical Music Theory.

- 2. ***State Sequence to MIDI Conversion***: This function takes the sequences of states generated by the agent, representing its composition, and translated them back into MIDI. This conversion was crucial for evaluating the agent's performance in a tangible and audible format, allowing us to listen to the music composed by the agent. It also facilitated the sharing and further processing of the generated compositions using standard music production software.

2.2 Learning Environment

In designing the learning environment for this music generation project, the primary objective was to manage the complexity of the state space while preserving the essential elements of musical composition. For preliminary explorations of the concept, the focus was specifically on monophonic compositions, where only one note is played at a time, which is a constraint that significantly reduces the overall state-space complexity.

The state space was constructed with the following dimensions:

- 1. ***Pitch Values***: The range of pitch values was limited to three octaves, providing a comprehensive, yet manageable set of notes. This constraint served to both mimic the typical range of most musical instruments that a composition could be voiced by (like a horn, guitar, or other stringed melodic instrument), and to significantly reduce the complexity of the state-space, while preserving the natural characteristics of most melodies to be within a reasonably local range of pitches.

- 2. ***Note Durations***: By discretely subdividing note durations into 8 separate values, the state space retains sufficient granularity to express rhythmic variations while keeping the state space simple.

- 3. ***Start Times***: To manage the temporal progression of compositions, 80 possible starting times for the onset of a note are included in this dimension, which correspond to 10 longest-duration whole notes played in succession. This discretization of start times allows for precise control over the temporal layout of the composition.

By focusing on these dimensions and omitting less critical features such as key pressure (which ranges from 0 to 127), we created a more tractable state space. The reduced state space is represented as a 3-D array, capturing pitch, duration, and start time, each within their defined ranges. With the preceding formulation, the state space consists of a total of $37 \times 8 \times 80 = 23,680$ possible states.

2.3 Actions

The action space was designed to align with the reduced state space, enabling the agent to make decisions that are both meaningful and computationally feasible. The action at each step consists of two components: selecting a new pitch value and determining the note duration.

- 1. ***Pitch Selection***: The agent can choose from the same range of pitches as defined in the state space, encompassing three octaves.

- 2. ***Note Duration***: Mirroring the state space, the agent has eight possible note durations to choose from.

This formulation of the action space results in a total of $37 \times 8 = 296$ possible actions, each representing a unique combination of pitch and duration. Furthermore, each action implicitly advances the start time of the next note, ensuring the temporal progression of the composition.

2.4 Rewards

The design of the reward system is crucial in guiding an agent to learn how to compose aesthetically pleasing melodies. For the reinforcement learning model, the methodology employed was a multifaceted approach to provide direction and encourage the generation of musically coherent compositions.

- 1. ***Use of Pre − Built Attention − Based LSTM Music Generator***: A key component of the approach was the integration of a pre-built attention-based LSTM music generator, developed by the Google Magenta project. This tool was particularly suitable for this project as it generates monophonic melodies within a 3-octave range. The output from this generator served as a 'goal' state, providing a high-quality, musically sound target for the agent to emulate.

- 2. ***Incorporation of Hard − Coded Musical Scales***: To impart some fundamental aspects of music theory, we included hard-coded musical scales within the learning environment. This feature allowed experimentation with teaching the agent basic principles of harmony and scale structures, further directing the agent's learning towards musically valid compositions.

- 3. ***Baseline and Positive Reward Structure***: The reward structure was designed to incentivize the agent towards desired outcomes. At each timestep, the agent receives a baseline reward of -1. This negative reward encourages the agent to actively seek better alternatives rather than remain inactive. If the agent selects a note that aligns with the key of the generated guide track, matches a note from the guide track, or uses a timing present in the guide track, it receives a positive reward. This reward system nudges the agent towards reproducing aspects of the guide track, aligning its actions with the goal states.

- 4. ***Goal State Representation and Reward Mechanism***: The conversion of MIDI to state sequences was instrumental in defining these goal states for the reinforcement learning model. By aligning the agent's actions with these goal states, we created a structured way to guide the learning process. The agent was encouraged to discover sequences of actions (notes and durations) that closely matched the target melody, thus learning to generate music that is harmonically and rhythmically coherent. This approach not only promotes the learning of musical structures but also fosters creativity within the confines of musically sound patterns.

By leveraging a combination of advanced LSTM-generated melodies, music theory fundamentals, and a nuanced reward structure, the agent is directed towards composing melodies that are not only technically sound but also aesthetically pleasing. This multi-tiered approach was pivotal in achieving the delicate balance between creativity and coherence in music composition.

2.5 Incorporating Human Feedback

In the process of training the model to generate musically coherent compositions, integrating human feedback proved to be a pivotal strategy. This approach allows the agent to refine the its learning trajectory based on subjective human evaluations, which are particularly important in the artistic domain of music composition. This strategy was inspired by Griffith's pivotal publication in 2013 [5].

During the training phases, periodic checkpoints have been established where the model's output, in the form of MIDI data, is played back for human evaluation. Listeners are asked to rate the composition on a scale from 1 to 10 across three distinct dimensions: creativity, naturalness, and authenticity. These dimensions were carefully chosen to correspond to key aspects of musical composition:

- ***Creativity***: Reflecting the novelty and originality of the composition.
- ***Naturalness***: Indicating how smoothly the notes and rhythms flowed together, akin to a human-composed piece.
- ***Authenticity***: Assessing the technical correctness of the piece, such as adherence to the correct key and harmonic structures.

The ratings collected from human feedback are directly aligned with the model's reward functions. For instance, a high rating in creativity influences the rewards associated with novel and unexpected note sequences. Similarly, ratings in naturalness and authenticity informs the rewards related to timing accuracy and key adherence. This alignment ensures that the model's learning objectives are continuously adjusted to align with human perceptions of what constitutes good music.

Based on the collected feedback, the reward values are dynamically adjusted the to reflect the user's suggestions. If certain aspects of the composition, such as timing or key, consistently received lower ratings, the reward structure was modified to place greater emphasis on improving these elements in subsequent training iterations. This dynamic adjustment allowed the model to focus on areas needing improvement while maintaining its strengths in other aspects.

The inclusion of human feedback bridges the gap between algorithmic learning and human artistic sensibilities. Incorporating human judgments into the training process ensures that the model's output not only adheres to technical musical standards but also better resonates with human listeners on an aesthetic level. The process of collecting feedback, adjusting rewards, and retraining forms a continuous loop. This approach fosters a collaborative environment where human input directly influences the model's development, leading to a more nuanced and appealing musical output.

2.6 Choice of Learning Control Approach: Off-Policy Monte Carlo

In the development of this music generation model, the selection of an appropriate learning control approach was crucial. After extensive experimentation and analysis, the decision was made to employ Off-Policy Monte Carlo with Importance Sampling. This choice was driven by several key factors, which are discussed in detail within this section.

The nature of this project, where each episode was guaranteed to terminate upon reaching the end of a music clip, made Off-Policy Monte Carlo a more natural fit compared to other methods. Unlike approaches like Q-Learning, which rely on continuous updates and can struggle with episodes that have definite endpoints, Monte Carlo methods are well-suited for tasks with clear termination points.

One of the primary reasons for choosing Off-Policy Monte Carlo was its avoidance of bootstrapping. Bootstrapping methods, such as Q-Learning and n-step TD, base their updates on other estimates, which can lead to inaccuracies, especially when future rewards are uncertain or difficult to predict. In the context of music generation, where future states (i.e., sequences of notes) can be highly variable and unpredictable, reliance on bootstrapping can lead to misguided reward estimations. Off-Policy Monte Carlo, on the other hand, estimates returns based on complete episodes, providing a more stable and reliable learning process.

The significant reduction of the state space in this project also influenced the choice of learning method. By ignoring note velocity, discretizing the time dimensions, and focusing on a pitch range of around 3 octaves, the truncated state space was compact enough to be effectively managed by a tabular approach. This reduction in complexity made the tabular representation feasible for Off-Policy Monte Carlo, allowing for efficient learning without the need for function approximation.

Prior to finalizing the choice, Q-Learning and n-step TD methods were also applied to the task. However, these methods exhibited notable issues related to bootstrapping. As the values of 'n' increased in n-step TD, we observed improvements in performance, but these were still overshadowed by the inherent instability of bootstrapping. These observations further reinforced the decision to focus on optimizing the performance of the Off-Policy Monte Carlo method for this music generation task.

Section 3: Initial Experimentation: Basic Scales

3.1 Setup

In the preliminary phase of this project, the primary objective was to evaluate the agent's capability to grasp fundamental concepts of music theory. This approach mirrors the pedagogical process in music education, where novices, typically children, are first introduced to playing a musical instrument through learning basic scales under the guidance of an instructor. To parallel this learning trajectory, the scale of B flat, encompassing three octaves, was chosen as the target state within the experimental environment.

For the initial setup, key parameters were established as follows: $\epsilon = 0.1$ and $\gamma = 1$. However, upon further contemplation regarding the nature of the goal states, it became apparent that the original γ value might cause an excessive focus on distant goal states, overshadowing the importance of proximal targets. This realization prompted an adjustment of the gamma parameter to 0.1. This modification led to significantly enhanced results, suggesting a more balanced approach in the agent's decision-making process, particularly in terms of near-term goal prioritization.

Maintaining a relatively low epsilon value was found to be conducive to the model's convergence towards a melody that exhibited greater musicality. It is important to acknowledge that assessments of musicality are inherently subjective. This aspect presents a compelling opportunity for future exploration and experimentation, potentially leading to further refinements in the model's parameters and its ability to replicate human-like musical comprehension and creativity.

3.2 Results

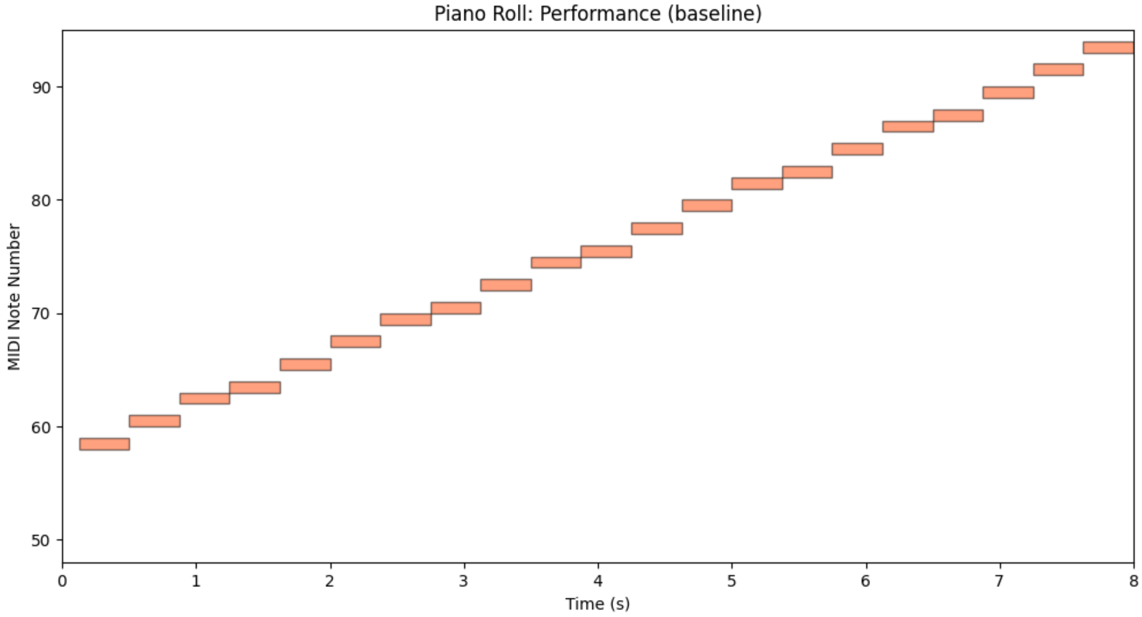

Figure 1: Bb Major Scale Visualization

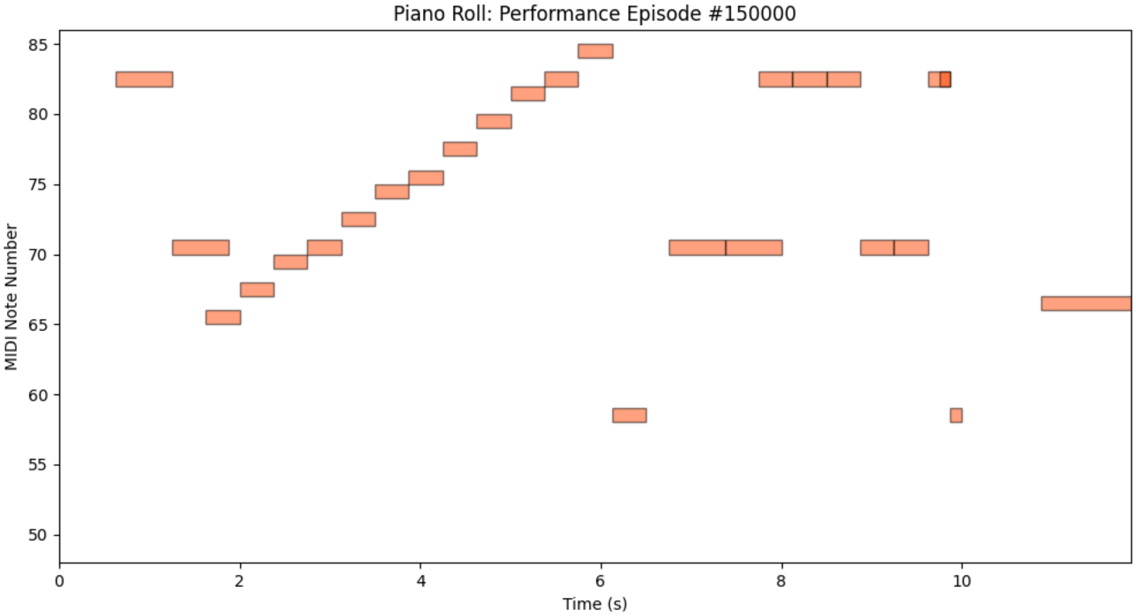

Figure 2: Final Performance Output (after training on Bb Major Scale)

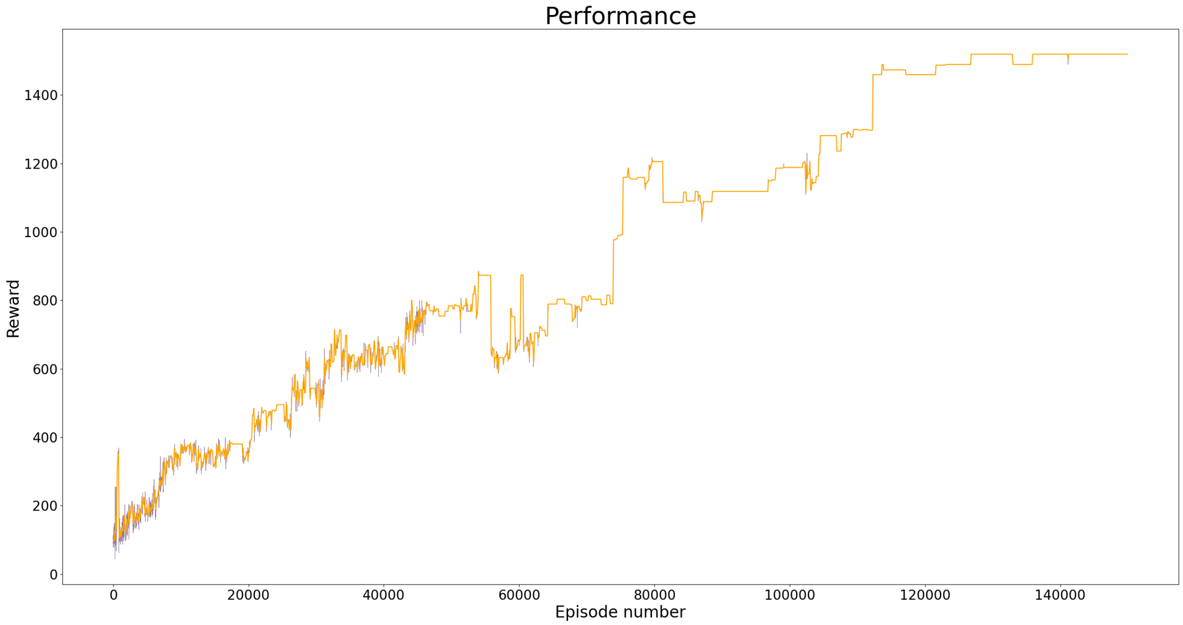

Figure 3: Agent Reward-based Performance over 150,000 episodes of training on Bb Major Scale.

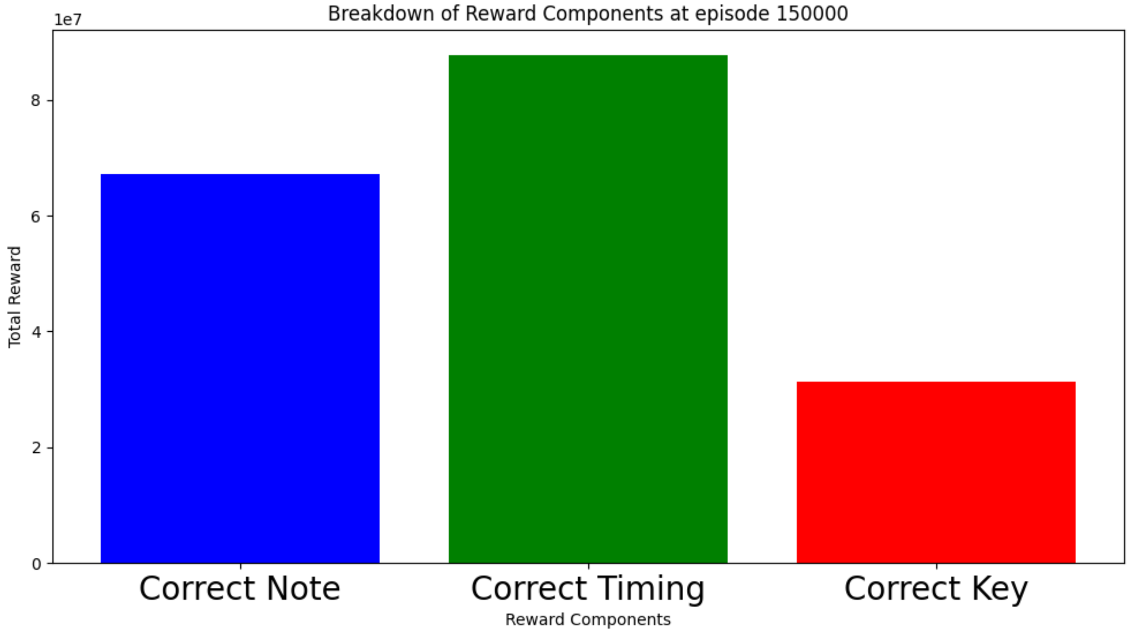

Figure 4: Comparison of Reward Signals received during 150,000 episodes of training on Bb Major Scale.

3.3 Discussion

The agent's ability to closely match the target B flat major scale is evident when comparing Figures 1 and 2. Over 150,000 episodes, the agent learned to play the scale with a high degree of accuracy, as shown by a quantitative analysis where 57.1% of notes played matched the target scale. This was achieved through the implementation of Off-Policy Monte Carlo Control combined with Importance Sampling and a strategically designed reward system. The adjustment of the gamma parameter from 1 to 0.1 was a pivotal change, reflecting the importance of immediate reward signals in the agent's learning trajectory, akin to immediate feedback in human skill acquisition.

The maintained low epsilon value allowed the agent to explore new sequences while still exploiting known rewarding states, facilitating convergence to a melody with enhanced musicality, a term which, despite its subjective nature, was operationally defined by criteria such as tonal accuracy and adherence to rhythmic structure.

The reward learning curve in Figure 3 displays a trajectory of steady improvement, with occasional regressions that may signify the agent's exploration of the musical space. The breakdown of reward components in Figure 4 reveals that correct timing was the most significant challenge for the agent, perhaps indicating a more complex temporal understanding required in music learning. Figure 4 also illuminates the fact that notes played that were technically incorrect in terms of the strict sequence, but were in the correct key, still appeared to be relevant in the total reward received by the agent. This helps to explain some of the note choices made that were not in the original sequence.

Section 4: Further Experimentation: Unique Melodies

4.1 Setup

In the next phase of this project, the primary objective was to evaluate the agent's capability to learn a unique melodic sequence. This approach was taken to illustrate the agent's ability to generalize its learning control process to any monophonic melody. To generate the sequence, a pre-trained attention-based LSTM music generator (courtesy of Google's Magenta project [3]) is used to generate new melodies. This model was designed to create monophonic melodies within a three octave range, making it a solid choice for generating the goal states for this experiment.

Both ϵ and γ were left unchanged from the previous experiment. This choice was made simply because the performance was reasonably solid in the initial experiment. In the future, it could be lucrative to experiment further with increasing the rate of exploration of the model.

4.2 Results

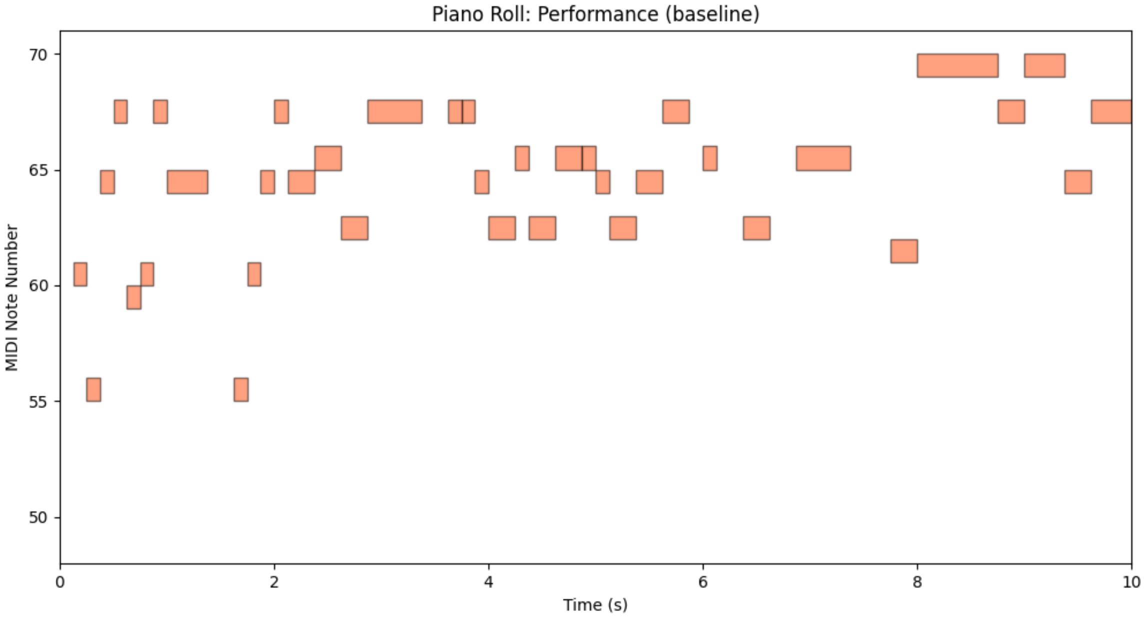

Figure 5: Unique Melody Visualization

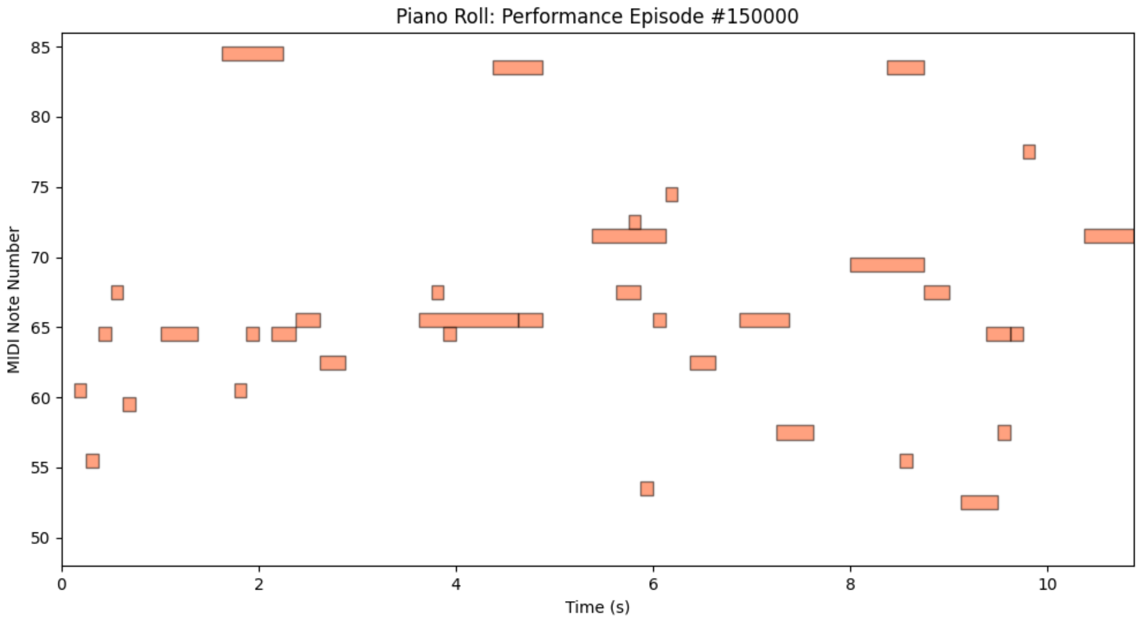

Figure 6: Final Performance Output (after training on Unique Melody)

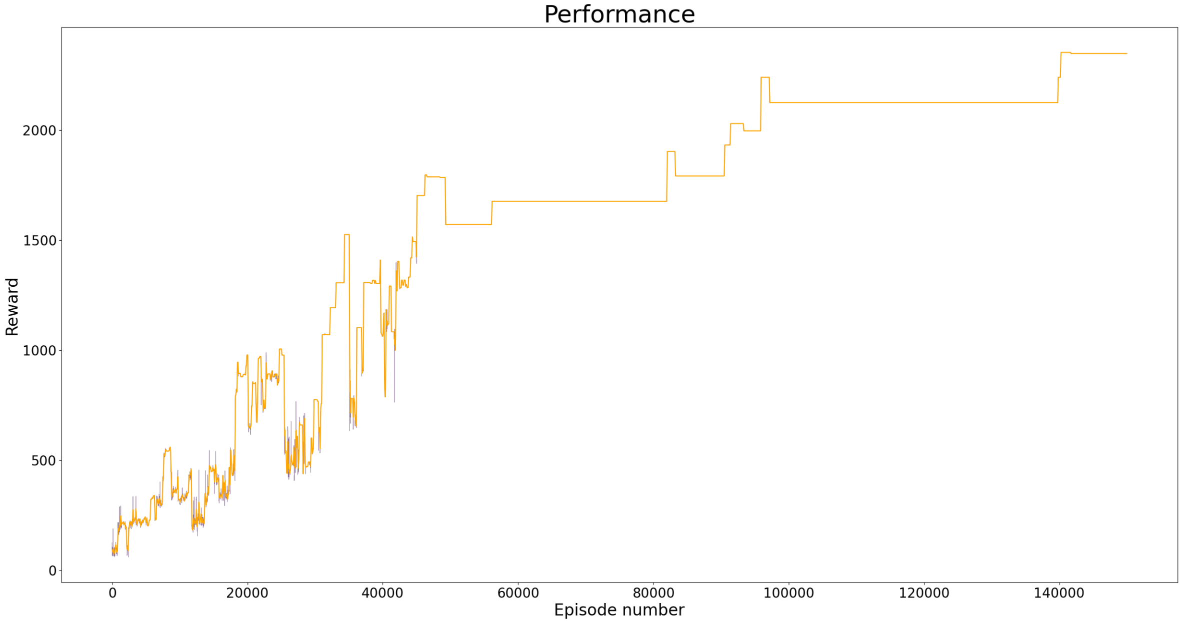

Figure 7: Agent Reward-based Performance over 150,000 episodes of training on Unique Melody.

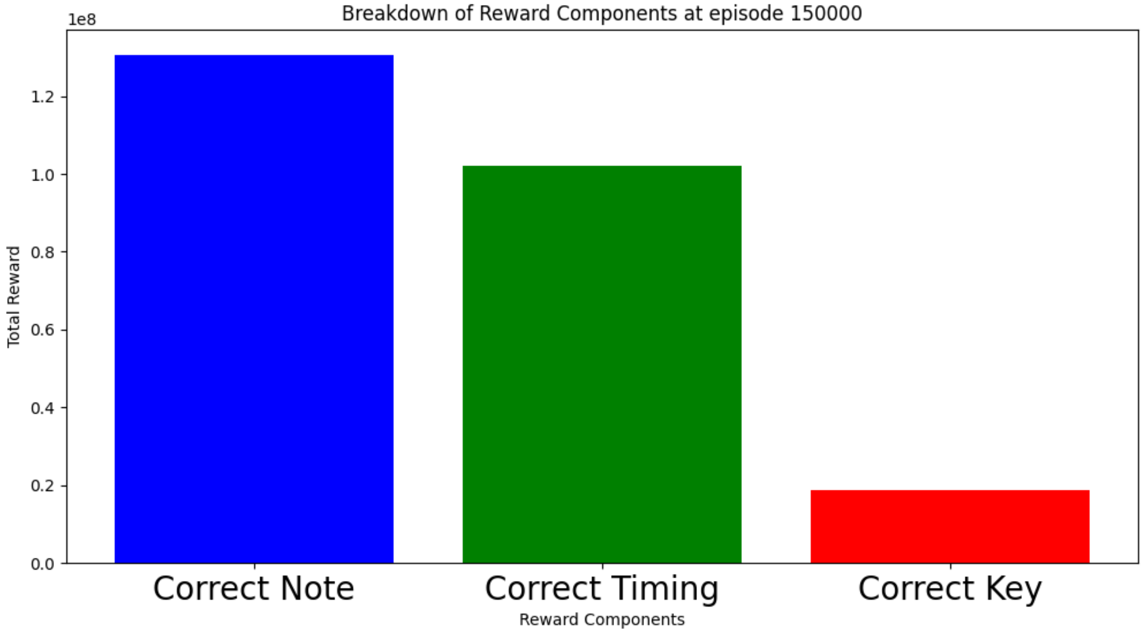

Figure 8: Comparison of Reward Signals received during 150,000 episodes of training on Unique Melody.

4.3 Discussion

In an analytical comparison of Figures 5 and 6, a discernible congruence emerges between the final piano roll performance rendered by the model and the baseline sequence generated via the pre-trained model. This similarity is quantitatively substantiated by the presence of 17 identical notes out of the 36-note baseline sequence in the model's learned rendition. This overlap signifies a notable degree of learning and adaptation within the model's algorithmic framework.

Further examination of Figure 7 reveals a critical phase in the model's learning trajectory. During this initial phase, characterized by significant noise and variance, the model exhibits a pattern of trial and error, indicative of an exploratory learning approach. This phase is pivotal in the model's development, as evidenced by the incremental improvements, quantified by the escalating rewards.

The data presented in Figure 8 provides insight into the reward distribution mechanism within the model. It indicates a predominant emphasis on the accuracy of note selection as the primary factor contributing to reward acquisition. Concurrently, there is a relatively lesser focus on timing accuracy, followed by a markedly reduced emphasis on adherence to the correct key. This distribution pattern of rewards suggests a hierarchical approach in the model's learning algorithm, prioritizing note accuracy over temporal precision and key fidelity.

Section 5: Final Experimentation

5.1 Setup

In the concluding phase of our experimentation, I began to embark on an exploration of a model adept in handling polyphonic sequences. The complexity inherent in polyphony necessitated an innovative approach in both the model's design and its operational parameters.

- 1. ***Modifications to the State and Action Spaces***: To facilitate this exploration, we introduced an additional dimension into the state and action spaces of the model. This modification was imperative to accommodate the intricate nature of polyphonic music. The new dimension specifically encodes the type of chord, a critical element in polyphonic composition

- 2. ***Encoding of Chord Types and Simplification Strategy***: In the realm of music theory, a myriad of chords can be voiced using a limited set of pitches. I narrowed the focus to eight commonly used chords in popular music, integrating these into the model's new dimension. Each added pitch, in the traditional sense, could represent up to 37 possible notes, thereby exponentially increasing the complexity. This strategic encoding enables the model to generate polyphonic sequences with relative complexity, bypassing the overwhelming intricacy that would arise from encoding additional pitches.

- 3. ***Revamping the Reward System***: A pivotal alteration in this phase was the overhaul of the reward system. To mitigate excessive repetition and promote diversity in the model's output, we introduced a decaying factor for consecutively playing the same chord. This was operationalized through a short-term histogram of recently played chords.

Under this system, a chord's reward diminishes according to its frequency in recent play, following the formula $6 \times \frac{1}{2^k}$, where k denotes the number of times the chord appears in the histogram. This nuanced reward mechanism proved instrumental in refining the model's final output.

5.2 Results

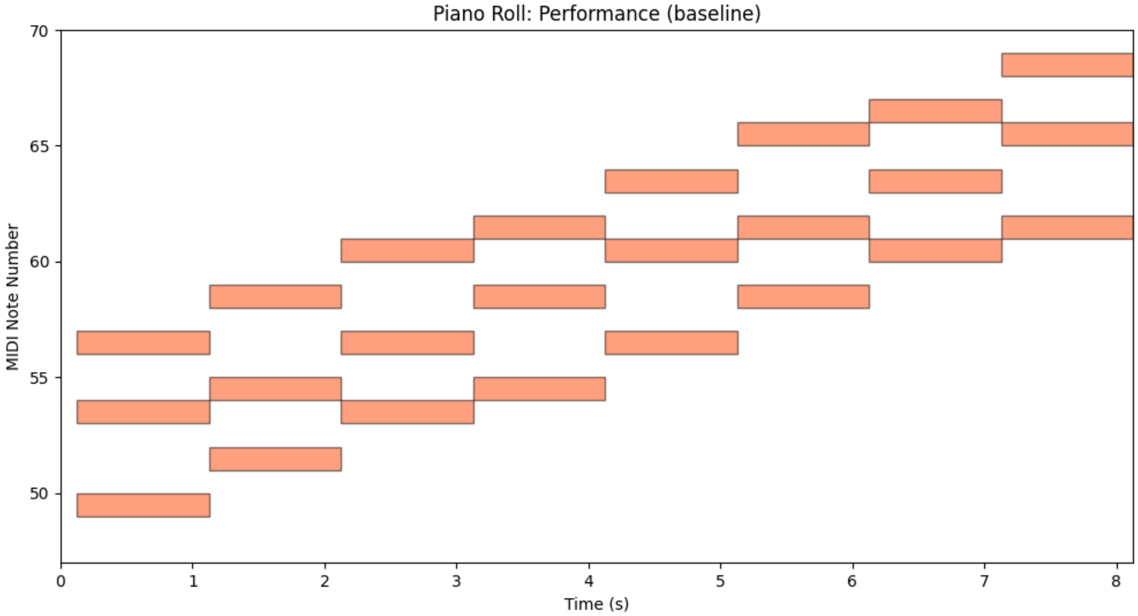

Figure 9: Chord Sequence Visualization

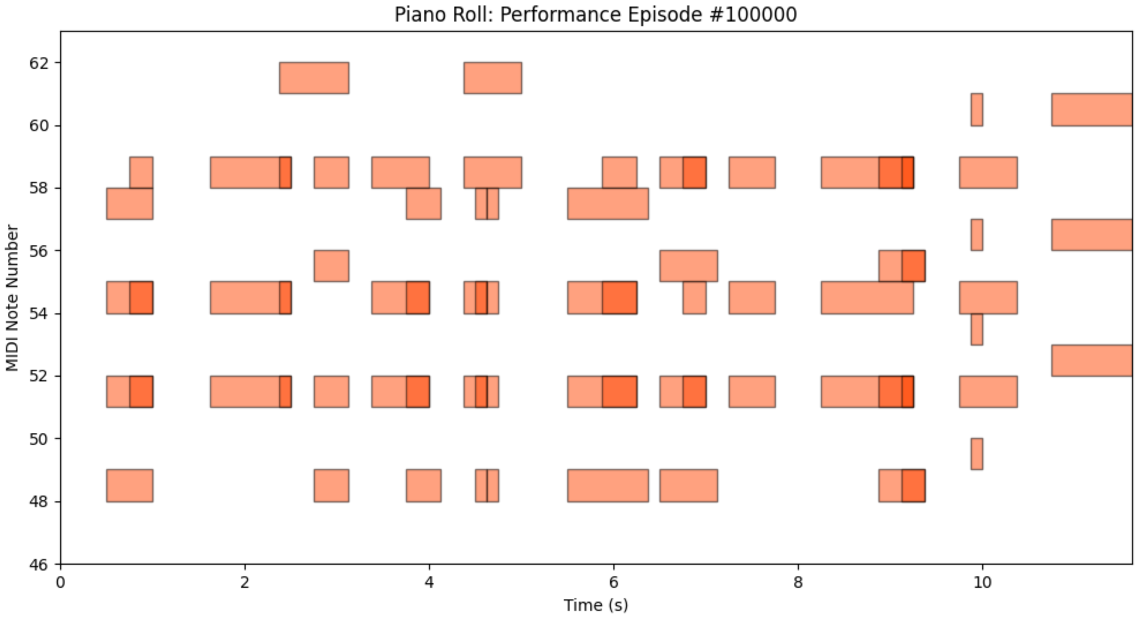

Figure 10: Final Performance Output (after training on Chord Sequence)

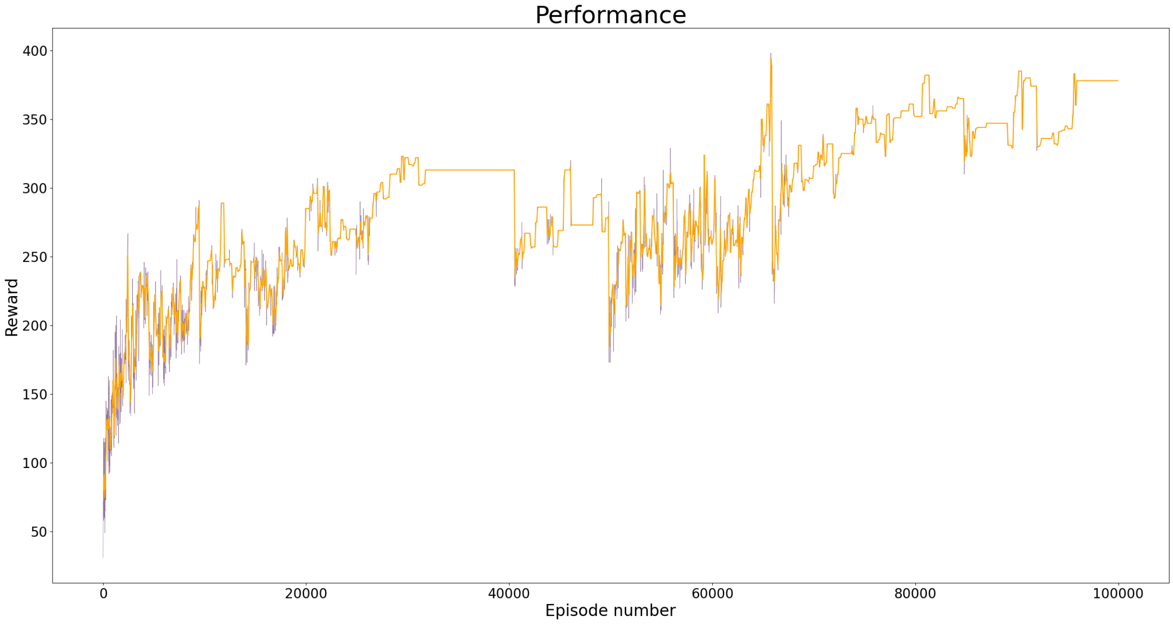

Figure 11: Agent Reward-based Performance over 150,000 episodes of training on Chord Sequence.

5.3 Discussion

The piano roll output depicted in Figure 10 reveals a series of musically intriguing choices. A notable observation is the model's tendency for a consistent alternation between chords. This pattern is likely attributable to the model's strategic optimization of its reward function. The agent appears to have developed a mechanism to maximize rewards through specific chord progressions, demonstrating a level of understanding of the musical structure within the confines of the programmed reward system.

An additional point of interest lies in the encoding of chords as enumerated values from 1 to 8. This method, while efficient, does not account for the nuanced relationships between similar chords, such as the proximity of a major triad to a major seven chord. Despite this lack of detailed encoding in the state space, the model exhibited an inclination to select these harmonically related chords, likely influenced by the reward system that incentivizes the inclusion of individual notes from the original sequence within the chords played.

A critical observation from this experiment is the model's disregard for the overarching ascending arc of the baseline sequence. This oversight can be directly linked to the design of the reward function, which did not explicitly value the preservation of the original sequence's macro-structural elements. The performance plot, which shows a consistent upward trend in total rewards over increasing episodes, does not shed light on this aspect of the model's behavior, indicating a potential area for refinement in future iterations of the model and its reward scheme.

An interesting avenue for future research might involve the expansion of the histogram size used in the reward function. By doing so, the repetition of recently played chords could be further minimized, potentially leading to a more varied and innovative musical output. Such adjustments could address the current limitations in the model's ability to diversify its chordal selections.

Section 6: Potential Future Improvements

6.1 Non-Tabular Methods

One of the primary areas for future development is the exploration of non-tabular methods. The current model utilizes a tabular approach due to the reduced state space. Expanding to non-tabular methods, such as function approximation techniques using neural networks, could be an interesting approach to enable the model to transcend the issue of neatly handling larger and more complex state spaces. This expansion would potentially allow for more expressive performances, with more variance in velocity and timing of musical phrases.

6.2 Experimental Reward Systems

Another area for improvement lies in the design of more sophisticated reward systems. Future iterations of the model could experiment with reward systems that impose certain musical constraints, such as limiting exploration to diatonic scales. By constraining the scale to the key of the song, the agent could be encouraged to focus more on timing and rhythmic complexity, potentially leading to more harmonically consistent and rhythmically interesting compositions. Incorporating additional feedback mechanisms that adapt the reward system based on the agent's performance could further refine its learning process, aligning it more closely with musical aesthetics.

6.3 Curriculum Learning

The implementation of Curriculum Learning could also significantly enhance the model's performance. In this approach, a 'teacher' model would set specific tasks for the 'student' agent, focusing on particular aspects of music composition such as pitch accuracy or timing. This targeted approach would allow the agent to concentrate on and improve specific weaknesses in its compositions. By breaking down the learning process into more manageable and focused tasks, the agent could achieve a more nuanced understanding and execution of complex musical elements. Additionally, this method could accelerate the learning process by providing structured guidance, leading to more rapid improvements in the quality of the generated music.

Section 7: Conclusion

In conclusion, this exploration into the realm of music generation using reinforcement learning has yielded both promising results and insightful directions for future experimentation. The series of experiments conducted, ranging from basic scale replication to the generation of unique melodies and the intricate handling of polyphonic sequences, have demonstrated the significant potential of reinforcement learning in capturing the essence of musical creativity and structure.

The initial experiments with basic scales laid the foundation, proving the model's capacity to assimilate fundamental musical concepts and replicate them with a high degree of accuracy. Moving forward, the experiments with unique melodies and polyphonic sequences showed the model's versatility and adaptability in dealing with increasingly complex musical structures. The introduction of novel elements, such as the encoding of chord types and the dynamic reward system, marked a significant advancement in our understanding of how reinforcement learning can be applied to music generation.

Throughout these experiments, the use of MIDI as the primary data format proved crucial. Its structured representation of musical elements provided the necessary framework for the reinforcement learning model to understand and generate music. The choice of Off-Policy Monte Carlo Control as the learning approach was pivotal, enabling the model to learn from complete episodes and avoid the inaccuracies associated with bootstrapping methods.

Incorporating human feedback into the training process was a key innovation, allowing for a fusion of algorithmic precision with human musical sensibilities. This feedback loop not only refined the model's learning trajectory but also ensured that the generated compositions resonated with human listeners on an aesthetic level.

As we look towards the future, several avenues for improvement and exploration stand out. The potential use of non-tabular methods, such as neural networks, promises to handle larger and more complex state spaces, enabling more nuanced musical generation. Experimenting with different reward systems and incorporating curriculum learning could also further enhance the model's ability to produce harmonically consistent and rhythmically captivating compositions.

References

[1] Textbook: Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction, Second Edition, MIT Press (2017).

[2] Lecture Videos: Professor Jivko Sinapov, Lectures Weeks 1-13, Tufts University.

[3] Google Magenta, *https://magenta.tensorflow.org*

[4] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[5] Griffith et al, "Policy shaping: Integrating human feedback with reinforcement learning", Advances in neural information processing systems (NeurIPS). 2013.

[6] A Generative Model for Creating Musical Rhythms with Deep Reinforcement Learning (2021): *https://aimc2021.iem.at/wp-content/uploads/2021/06/AIMC_2021_Karbasi_et_al.pdf*

[7] Bach2Bach: Generating music using a Deep Reinforcement Learning approach (2018): *https://arxiv.org/ftp/arxiv/papers/1812/1812.01060.pdf*

[8] A survey on Deep Reinforcement Learning for Audio-Based applications (2021): *https://arxiv.org/pdf/2101.00240.pdf*

[9] F. Shah, T. Naik and N. Vyas, "LSTM Based Music Generation," 2019 International Conference on Machine Learning and Data Engineering (iCMLDE), Taipei, Taiwan, 2019, pp. 48-53, doi: 10.1109/iCMLDE49015.2019.00020.

[10] David's MIDI Spec, *https://www.cs.cmu.edu/ music/cmsip/readings/davids-midi-spec.htm*