CSC413: Seamless Song Transitions via Recurrent Neural Networks

Feiran Huang Haike Yu Suqing Liu Yo-Hsuan Chen
Department of Mathematical and Computational Sciences
University of Toronto Mississauga
{philipp.huang, haike.yu, suqing.liu, yohsuan.chen}@mail.utoronto.ca

Abstract

This paper introduces SynthWave Maestro, an innovative AI-driven DJ application that employs Recurrent Neural Networks (RNNs) for seamless song transitions. The core functionality of this application is built on RNNs, which analyze and merge audio data from diverse songs to create uninterrupted playback experiences. Our primary focus is on audio-to-audio generation, aimed at redefining traditional DJing practices. The application is designed to recognize and blend the rhythmic and structural elements of different tracks, thereby facilitating the generation of novel audio content.

The dataset used in this project encompasses a wide range of genres, tempos, and complexities, providing the RNNs with extensive exposure to varied musical styles. This diversity ensures the model's adaptability in creating harmonious transitions across distinct tracks. The paper details the training process of the RNNs, their capabilities in understanding musical intricacies, and the methodologies employed in generating smooth connections between songs. Ultimately, SynthWave Maestro represents a step forward in augmenting the art of DJing, offering a unique tool for DJs to craft immersive and continuous musical experiences.

1 Introduction

The evolution of music, combined with the advancements in technology, has continually shaped and redefined the way we experience and interact with audio compositions. In recent years, the fusion of artificial intelligence and music has led to groundbreaking innovations in the field, pushing the boundaries of what was once considered achievable. One such innovation lies in the convergence of AI and the art of DJing, where technology moves beyond mere playback to become an active participant in the creation and curation of musical experiences.

In recent years, various AI-music combination applications such as SongMASS[1], AudioCraft[2], and DeepRapper[3] have ignited a revolution in the music industry, showcasing the tremendous potential of artificial intelligence in reshaping how we compose, produce, and consume music. However, while these applications proficiently handle text-to-audio and audio-to-text conversions, our project ventures into uncharted territory by pioneering audio-to-audio generation, a domain yet unexplored in the realm of AI-driven music applications.

We present SynthWave Maestro, an innovative AI DJ application that harnesses the power of Recurrent Neural Networks (RNNs) to facilitate seamless transitions between songs, fundamentally transforming the conventional DJing landscape. This groundbreaking endeavor marks a significant leap forward by enabling the generation of new audio content based on existing musical pieces, fostering an unprecedented level of continuity and cohesion between disparate tracks.

This groundbreaking AI DJ project not only represents an advancement in technology but also signifies a paradigm shift in the way we conceive, produce, and experience music. The fusion of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

artificial intelligence and DJing not only augments the capabilities of DJs but also democratizes the art form, offering a potential for broader accessibility and innovation in music curation and redefining the very essence of musical expression and engagement.

2 Background and Related Work

2.1 Background of DJing

Music composition, especially in the realm of DJing, is a complex art that asks for a deep understanding of musical structure and theory. A variety of factors comes into play when determining whether a piece of audio resonates harmoniously—ranging from the pitch of each note and the intervals between them to the duration and intensity of each note. The art of DJing primarily revolves around transitioning smoothly from one track to another, maintaining a consistent and engaging auditory experience for the audience.

The emergence of machine learning, particularly deep learning techniques, offers a groundbreaking approach to this challenge. This project specifically targets the automation of this aspect of music composition, focusing on the development of an AI model capable of analyzing and mimicking the intricate process of creating musical transitions.

2.2 Related work

Existing music generators akin to our project fall into two categories. One type concentrates on generating entirely new songs or extending existing ones, exemplified by Microsoft's Muzic [4] and some notebooks from kaggle, such as Music_generator_using_LSTM [5]. The second type specializes in extracting elements from two songs to blend them, as demonstrated by Spindrop [6]. While the fundamental goal of both these generators and our project remains consistent—creating a fresh musical piece from existing ones. The key distinction is that unlike these common music generators, our project places a primary emphasis on the seamless transition between two musical pieces. For in-depth information regarding Muzic and Spindrop, please refer to the attached references.

3 Data

This project revolves around the processing of two musical compositions, taking them in as input, and generating a transitional piece that bridges the two. In line with this objective, our team opted for the utilization of the Lakh MIDI dataset.

The Lakh MIDI dataset encompasses an extensive collection of 176,581 individual MIDI files, with 45,129 of them meticulously aligned and harmonized with entries in the Million Song Dataset. The principal aim of this dataset is to facilitate seamless large-scale retrieval of music-related information. This encompasses both symbolic retrieval, which exclusively employs the MIDI files, and audio content-based retrieval, where insights extracted from the MIDI files serve as annotations for their corresponding audio counterparts [7]. For our endeavor, we will employ the refined version of this dataset, which is made available on Kaggle [8].

4 Model Architecture and Figure

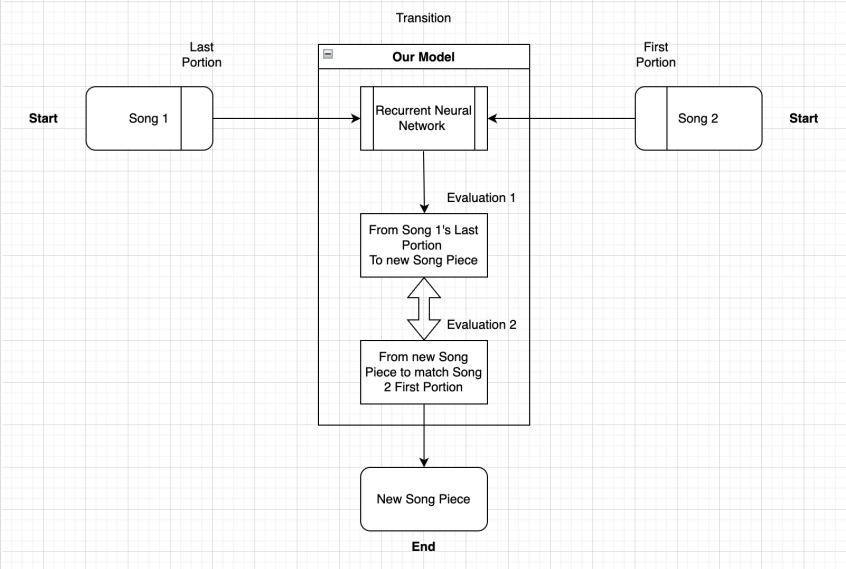

Figure 1: Architecture Figure

The core of our architecture involves feeding the RNN with a concatenated sequence comprising the concluding part of the first song and the beginning of the second song. The RNN, trained on a diverse range of MIDI data, is tasked with generating a middle piece that harmoniously bridges these two segments. This integrated approach allows the model to process the transition as a continuous sequence, ensuring a more natural and cohesive blending of musical elements.

The RNN must create a transition that maintains the integrity and flow of the first and second songs. The challenge lies in synthesizing a middle piece that respects the unique identities of both songs while providing a seamless auditory journey.

Our use of RNN leverages its strength in handling sequential data and its ability to capture the nuanced dynamics of music composition. The model's learning from MIDI data is crucial, as it provides an understanding of various musical styles and structures, enabling it to generate transitions that are both creative and contextually appropriate.

The evaluation mechanism is a critical component, assessing the coherence and musicality of the entire concatenated sequence, focusing on the newly generated middle piece. This ensures that the transitions sound natural and enhance the overall listening experience.

4.1 Method 1: 2 Song Generalization

Take in the first and second songs simultaneously, and cut the length of the longer song to match the shorter song.

Each note has four features: velocity (How loud the note was pressed), step (the flow of music), pitch, and duration.

Generate two labels at each step, one for each song.

Stop generating when the pitch is the same for some amount of time.

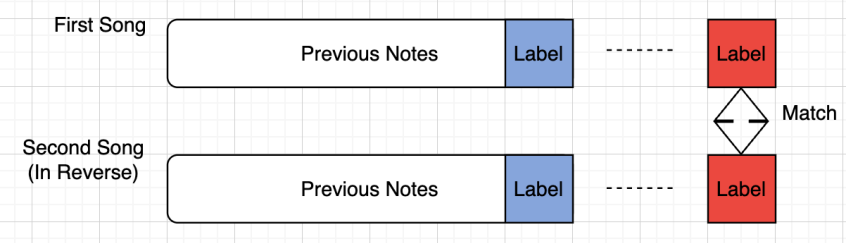

Figure 2: method1 Figure

4.2 Method 2: Same Song Generalization

This Idea utilizes the fact that a song should have a continuous style across; thus, training the RNN to regenerate the middle part of a song can make the RNN learn style.

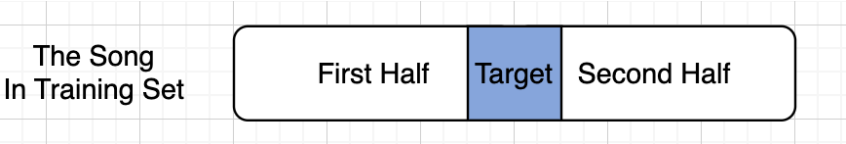

Figure 3: method2 Figure

Here we assumed if the rnn can reproduce a good middle part for one song then it has the potential to connect two different songs.

In the Training set, we choose the middle part of each song to be the target of that song. We then train the RNN to match the data in the training set.

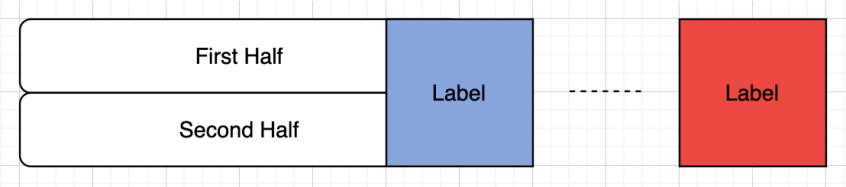

Figure 4: method2 Figure

5 Results

Owing to the training complexities associated with Method 2, we have decided to proceed exclusively with Method 1. The primary challenge with Method 2 stemmed from creating a custom dataset comprising pairs of songs and their seamless transitions. As such a dataset was not readily available online, and considering our limited resources to construct it independently, we are compelled to move forward with the more feasible approach offered by Method 1

5.1 Performance Metric

Training Set and Training Accuracy:

We train the RNN using pairs of songs, testing their training accuracy with the same two songs used in each training session.

Validation Set and Validation Accuracy:

The validation accuracy is assessed using different segments of the same two songs, ensuring that the model is evaluated on varied fractions of these tracks.

Human Feedback:

We gathered feedback by asking our friends and ourselves to evaluate the quality of the music generated by our model. The categories are Overall Flow, Overall Trend, and Length.

5.2 Quantitative Performance

Training Loss:

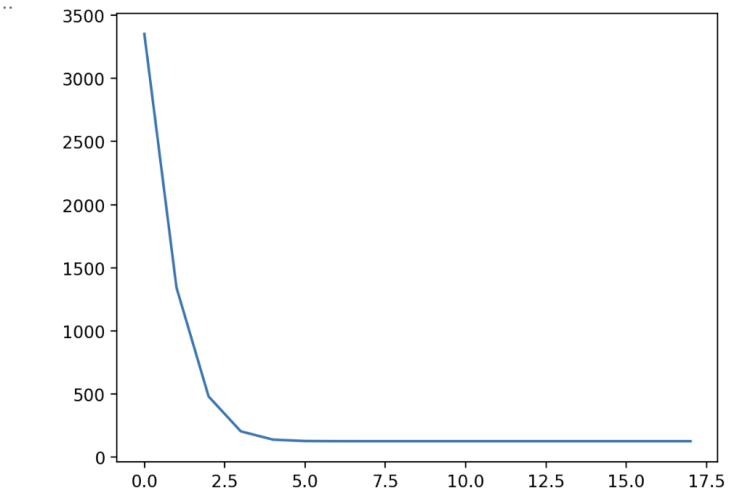

Figure 5: Training Loss Figure

Validation Loss:

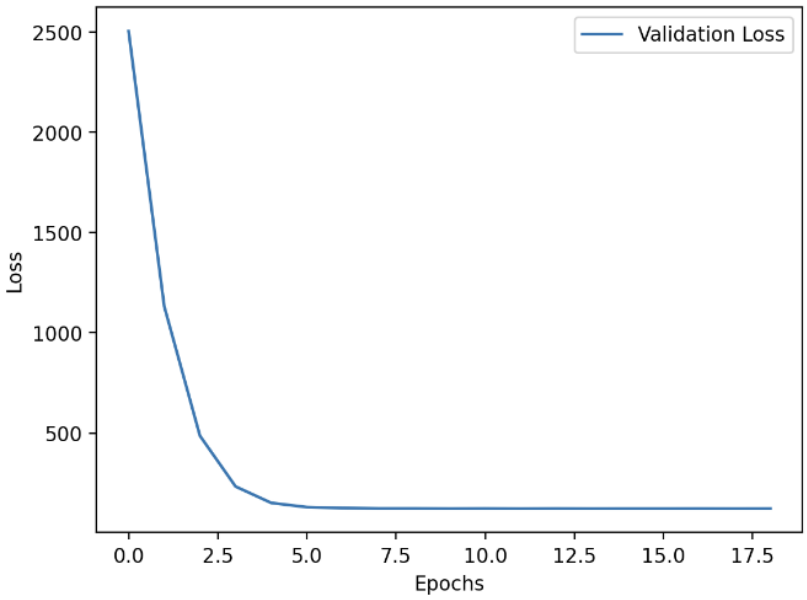

Figure 6: Validation Loss Figure

While the training and validation losses appear to be at a low level, it is indicative that the model may tend to be overfitting. This occurs due to the similarity in styles between the different sections of

the training and validation datasets, both from the same songs. However, we consider this to be a positive indicator. Our rationale is that since the model will be retrained for every new pair of songs, its tendency to overfit suggests a strong understanding of these songs. This capability is crucial for creating high-quality transitions between the two tracks. Thus, in this context, the model's overfitting is an asset rather than a drawback, aligning with our goal of generating seamless musical transitions.

5.3 Qualitative Performance

Human Feedback:

During our evaluation of the model's performance, we observed a tendency for the model to generate new notes until it identifies the optimal match continuously. This process can result in the creation of excessively long transitions, with up to thousands of new notes and lengths extending to as much as 8 minutes. Consequently, we have designated this iteration as the 'base model'. Moving forward, we have implemented a parameter in the new model to cap the maximum length of the generated transition, thereby addressing this issue. This is the 'new model'.

The evaluation will be out of 7 since we had 7 people in total, and each person will vote whether they liked or disliked the output based on the following metrics.

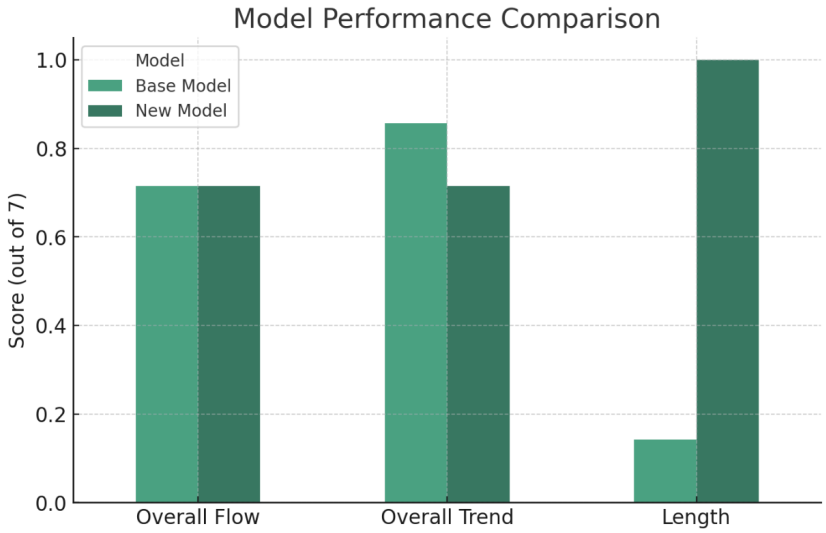

Figure 7: Human Evaluation Figure

The new model has shown significant improvement compared to the base model in terms of shortening the length of the newly generated piece. However, it did make a slight sacrifice in the Overall Trend. Fortunately, this sacrifice is almost imperceptible to our ears. When compared to the lengthy base model, the new model offers a reasonable length and generated sequence, thus proving to be superior to the previous model.

6 Discussion

While this new model appears promising on paper, it falls short of our initial expectations. There have been instances where the generated music does not meet our standards of quality. We believe that despite the model's high accuracy, it tends to train itself to replicate the characteristics of the training data without necessarily understanding music theory. It mimics existing songs but lacks the ability to grasp the underlying principles of music theory. Unfortunately, we have yet to determine how to instill this musical knowledge in the model. Several factors contribute to this challenge, with one of the most significant being our model's design, which we will delve into further in the next section.

7 Limitations

When designing our model, we encountered a challenge in training with multiple pairs of two songs. Consequently, we currently need to retrain the model each time we select a new pair of songs to generate the middle sequence. This presents an issue as it results in the model becoming overly specialized in the two trained songs, hindering its ability to learn beyond these specific tracks. Furthermore, it means that we have not effectively leveraged the diversity within our dataset.

This limitation arises from our limited understanding of training an RNN with sequential data such as songs. We believe that by exposing the model to a wider range of songs, it can gradually begin to generalize meaning and develop a deeper understanding of music theory.

Another notable challenge we've encountered is that the generated outcomes lack a pleasing auditory experience. Typically, songs tend to decelerate and fade out towards the end. This characteristic inadvertently influences our model, leading to newly generated sequences that often emulate this slowing, fading style. This pattern persists until the generated content transitions to elements inspired by the beginning of the second song.

8 Ethical Considerations

8.1 Intellectual property and copyright infringement

One of the most immediate ethical concerns is the potential for copyright infringement. The model may inadvertently create music that is too similar to existing copyrighted works, leading to legal issues. To mitigate this, the model must be designed to ensure originality in the transitions it creates, possibly by transforming the input sufficiently to avoid directly copying any copyrighted material.

8.2 Cultural appropriation

Music is deeply rooted in cultural expression. An AI model that borrows elements from music pieces without understanding their cultural significance can contribute to cultural appropriation, stripping elements from their original context and potentially disrespecting their cultural origins.

8.3 Psychological and emotional impact

Music has a profound impact on emotions and mental states. A model that generates transitions without understanding the emotional context could create jarring or unsettling experiences for listeners. It's important to consider how these transitions might affect listeners, particularly in contexts like therapy where music is used to support mental health.

8.4 Financial implications for the music sector

The introduction of this model has the potential to significantly alter established economic patterns within the music industry. The growing popularity of AI-crafted transitions may diminish the market for compositions created by human musicians, thereby affecting the traditional economic models for music production and distribution.

9 Conclusion

The development of SynthWave Maestro marks an advancement in the realm of AI-driven DJ applications. Through the integration of Recurrent Neural Networks (RNNs), this innovative tool redefines conventional DJing practices by enabling seamless transitions between diverse musical tracks. The application's core functionality, rooted in the analysis and merging of audio data, empowers DJs to create uninterrupted and novel playback experiences. Although we achieved promising validation and test accuracies, our pursuit of generating enjoyable music fell short of expectations.

However, we also believe there are many areas we haven't explored during the course of the project. We recognize that solely inputting digital music data into neural networks without integrating music

theory might not suffice for the generation of enjoyable music for human listeners. Recognizing the importance of music theory as a fundamental building block in the creation of harmonious and engaging music, we acknowledge its absence in our project and anticipate a more holistic approach that could potentially bridge the gap between technical accuracy and human-centric musical aesthetics.

In conclusion, SynthWave Maestro marks an exciting advancement in using RNNs for DJ purposes. It highlights the complex task of balancing technical accuracies with human-centric musical aesthetics. The path ahead involves improving AI-generated music to not just understand musical details but also connect well with how people perceive and enjoy music.

References

[1] Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. Songmass: Automatic song writing with pre-training and alignment constraint. *arXiv preprint arXiv:2012.05168*, 2020.

[2] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[3] Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. Deeprapper: Neural rap generation with rhyme and rhythm modeling. *arXiv preprint arXiv:2107.01875*, 2021.

[4] Microsoft. Muzic: A multi-modal deep learning framework for music generation. `https://github.com/microsoft/muzic`, 2023. Accessed: November 8, 2023.

[5] Abhishek Singh. Music generator using lstm, 2021. Accessed: November 8, 2023.

[6] Spindrop. Spindrop: Your source for music blending, 2023. Accessed: November 8, 2023.

[7] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, 2016.

[8] Imsparsh. Lakh midi clean dataset, 2023. Accessed: November 8, 2023.