**Proceedings of 12th IOE Graduate Conference**
***Peer Reviewed***
**ISSN: 2350-8914 (Online), 2350-8906 (Print)**
**Year: 2022 Month: October Volume: 12**

# Digital Music Generation using a Character-Level LSTM

Anil Pudasaini [a], Shashidhar Ram Joshi [b], Basanta Joshi [c], Sanjivan Satyal [d]

[a, b, c, d] *Department of Electronics & Computer Engineering, IOE, Tribhuvan University, Nepal*

[a] 076msice003.anil@pcampus.edu.np, [b] srjoshi@ioe.edu.np,
[c] basanta@ioe.edu.np, [d] sanziwan.satyal@pcampus.edu.np

**Abstract**

Deep learning for sequence modeling has gained prominence during the past few years. To accomplish this, LSTM network architectures have shown to be quite helpful for forecasting the next output in a series. There are several ways to generate music and this paper presents a novel way to generate digital music. Leveraging the power of sequence modelling, a novel variant of LSTMs, character level LSTM (Long Short Term Memory) is used to generate an adequate music. The created music is then thoroughly evaluated using a Turing test and an online musical audition. The model passed the Turing test and is able to generate music with good musical qualities making it challenging for the average listener to tell the difference between human composition and those produced by AI. The thesis provides a brief synopsis of the intuition, theory, and application of LSTMs in music generation, develop and present the network that is found to best achieve this goal, identify and include potential future network improvements.

**Keywords**

MIDI, music generation, LSTM, AI music, sequence modelling

## 1. Introduction

Music is a way of expressing emotions. The first thing to understand is that music is a language. It is an expression of feelings and emotions, but it is also a way for us humans to communicate with each other. Music has been around since the beginning of time. We can trace back the origins of music by looking at cave paintings and ancient carvings on rocks. We know that there were people who sang, played instruments, and danced long before we had written records or recorded sound waves on tape or vinyl discs. Music is used in religious ceremonies as well as celebrations such as weddings and funerals and just for fun in majority of the cases.

The creative process is an integral part of music composition. It involves the following steps: The initial idea or theme, which may be a fragment of a melody or a chord progression. This idea will guide musicians in the rest of the composing process. Musicians need to know what the musical piece will sound like before starting to write it down on paper. If they don't have an idea for the composition, then they just listen to some existing pieces and try to copy them as closely as possible. Once the musicians have an idea, they start working on it by playing around with chords and progression and try and improvise until a good sound music is generated.

Deep learning is a science of mimicking human thought process and in recent years we can see it being used in almost every field: from recommending a movie or a song to helping astronauts to perform works that require great exertion in space. It has been thought that creating music is far from the horizon of machines and this remained a fact until recent years. Now with the advent of deep learning algorithms, the dawn of music composition is here. MIDI is a standard developed by Dave Smith in 1983 for connecting equipment. No sound is sent via MIDI, only digital signals which are also known as event messages which instructs pieces of instruments. In this research, I have tried to implement a deep learning algorithm to showcase a way of music composition which was thought impossible some years back. The basic idea is to convert a MIDI music into a CSV file format and extract essential features in the form of sequence of characters from the file and then feed it to the neural network which then trains on the input data and then produces similar sequence of characters.These characters can now be reverse engineered to produce MIDI music.

## 2. Problem Statement

With the advancement in computer technologies and machine learning, increased use of machine learning algorithm can be seen in almost every fields. Despite the fact that there are several ways for artificially creating music, architectural design is difficult, training parameters are poorly understood, and generalization error and confusion in the most suitable method to model complicated data persist. By considering such issues, an approach to convert the music data into a suitable format and then train the model to generate music is a must. Similarly, instead on using only a single track to train the model as stated in [1], a model that can train on more tracks is needed. The problem of music creation is because of the fact that to be able to write or compose a music one needs to understand music theories and such. Use of AI to generate music of certain style can be revolutionary in the music industry. Using AI, totally original music can be generated. Hence, It can be a helpful creativity tool for even a professional artist. And in recent years there have been several issues due to copyright in Youtube background music as the trend of making Youtube videos is increasing. Because of this, there is a severe need for a copyright free music generation system. With the help of the this research, a hope to solve these issue is at the horizon.

## 3. Related Work

With the progression of time, technologies have reached zenith. In the field of artificial intelligence, numerous possibilities have emerged in the creation, synthesis, manipulation and enhancement of music with the aid of computer systems.

The earliest known recording of computer synthesized music was in 1951 A.D, produced by Alan Turing's Computing Machine Laboratory. It used the concept of algorithmic audio composition. This aproach of music creation hinted numerous opportunities in the field of intelligent music creation. Iannis Xenakis, a musician and engineer, turned into some other pioneer on this subject, the usage of stochastic possibilities to assist him create music. A stochastic procedure is a mechanism having randomly dispensed chance distributions that can not be expected however can be statistically studied. In the early 1960s, he used computer systems and the FORTRAN programming language to interweave numerous chance features to decide the overall shape of a composition in addition to several different factors (inclusive of pitch and dynamics) [2].

After this, instead of generating music from simple algorithm, generative modelling was more focused. David Cope was a professor of music in the 1980s, with his Experiments in Music Intelligence (EMI) [3], that the insight of computer composition might encompass a better knowledge of music via his three primary methods:

1. Signatures (commonalities – keeping what signifies style.
2. Compatibility (recombinancy – reuniting musical elements into new works.
3. Deconstruction (separation and analysis into parts)

Looking at [4] for comprehensive survey on deep learning based music generation, recent trend is on the composition of music using deep neural networks due to the generality as opposed to systems as grammar-based [5] or rule based music creation architectures [6]. Different music composition requires combination of different model in unison. In addition, a deep learning model if supplied with a large scale music dataset can learn the musical styles from the musical corpus in a subtle way and then generate new audio content. According to Fiebrink and Caramiaux [7], few advantages are :

1. If the required application is too complex to be specified in an analytic or manual brute force design, it can be built using a deep learning algorithm.
2. Algorithms that learn are oftentimes less brittle than manually written rule sets, and learnt rules are more likely to adapt properly to new situations with variable inputs.

Furthermore, unlike organized expressions such as rules and grammars, deep learning is unique in that the layer handles unstructured input that produces high-level expressions tailored to the task.

With the fact that deep learning algorithm are the de-facto methods for music composition of the recent age, in this research, possible use case of a much familiar feed-forward neural network, recurrent neural

network is explored. Music composition is a form of sequential modeling task. As we can see today, sentence completion tools like in Emails make use of deep learning algorithms such as RNNs and its variants to auto complete a sentence. In similar fashion, this research aims to make use of such technique to generate a sequence of texts which can later be converted back to music. Referring to [4] again, for the composition of melody, a sequential system using RNN and iterative feedforward sampling seems an approriate choice to go with. Sequence of texts processing are generally being used in Natural Language Processing (NLP) and this very concept can also be applied to monophonic music creation with the availability of new music manipulation tools.

For a limited sequences, a single input sample is processed by a RNN at each timestep resulting a trained network after the training phase is over. Using an algorithm that helps in reducing error across timestep by backpropagating error in time, a sufficiently trained network can be achieved however for a long sequence, backpropagated gradients can be null resulting a saturation state for learning.Also known by the name vanishing gradient problem. This problem can be eased using Long short-term memory (LSTM) networks which makes use of specialized gates [8]. In the last few years, an increased attraction towards computer aided music generation is being observed, Magenta research project by famous gaint company Google [9] was initiated. Futhermore Spotifys invention 'Creator Technology Research Lab' (CTRL) added a tool that helps artist in their creative process [10]. An approach based on LSTM architecture was recently published that uses single track to train and produce a MIDI song using a Music21 library's converter. Using the prebuilt libraries to parse information, a network is defined that predicts the note and duration [1]. Research on various approaches such as generic sequence modeling, conditional natural language processing,symbolic music generation are dominantly active today.[11]

## 4. Methodology

### 4.1 Dataset and Libraries

For the model to generate a good music, a dataset plays an important role. Generally combination of neural networks as well as deep learning architectures tend to require tremendous amounts of examples to function properly so a dataset should be sufficient size for learning to be accurate. "I feel that this tradeoff between the size of a dataset and its coherence is one of the primary difficulties when creating deep generative models. A decent generative model should be able to discriminate between distinct subcategories and generalize well if the dataset is exceedingly varied. If there are only minor changes between subcategories, however, it is critical to determine if the "averaged model" can provide musically interesting outcomes.", as said by Hadjeres in [12]. Unlike in the image processing domain where there are standard datasets such as MNIST dataset [13],there are none yet in the music domain. Finding a proper dataset for all the processing,analysing task is an issue. So the best option is to look for numerous libraries and datasets available online and that are open source. A few of them are listed below:

- Yamaha e-Piano Competition dataset
- The Classical piano MIDI database
- JSBach midi dataset
- BitMidi dataset with 113,245 MIDI files curated by volunteers around the world

### 4.2 Requirement gathering and Workflow

Regarding the dataset for the intial phase of the research, as there is no standard dataset for reference, manual searching and downloading the appropriate midi songs and preprocessing was carried out. In the requirements gathering phase,several websites for midi music were searched. In addition to that, converters such as csv-midi and midi-csv that are available on the fourmilab site were mustered. Every CSV representation of a MIDI file contains at least three fields namely:Track, Time and Type. Track contains information about the track to which record it belongs. Time contains information about different events takes place and Type contains information that identifies the type of the record. In addition to it, different file structure records such as : (0,0, Header), (Track,0,Start_track) etc convey meaningful information.

Furthermore File Meta-Events points to different events that take place with MIDI tracks. Channel Events are the main events in consideration as the variation of information inside them is what actually direct different instruments to play the music.Inside

the CSV files all the channel events have a suffix "C" and are of the form: (Track, Time, Note_ON_C, Channel, Note, Velocity),( Track, Time, Note_OFF_C, Channel, Note, Velocity) with each item having specific meaning. Initially the midi music will be preprocessed using a MidiEditor software. Essentially, what I did was removing snares,bass, etc to simply the song so as to utilize a single main channel of each song. After this, the conversion of MIDI to CSV is conducted utilizing the midicsv program. Again, further processing is done primarily to transform the CSV into a form of acceptable vector to LSTM.

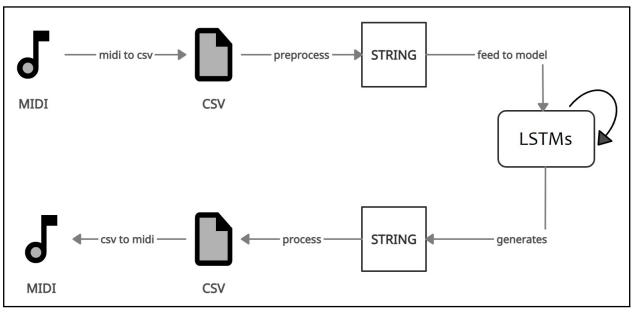


**Figure 1:** Steps in Music Generation

### 4.3 Model Intuition and Architecture

**RNN :** Recurrent Neural Networks are a special type of looped neural networks where the output from a layer is fed back to the layer to realize the output. There are input layers, hidden layers and output layers. At any given time t, the current input is dependent upon the input of current time and a previous time.

RNNs were created as there were problems with feed-forward neural networks in tasks such as sequential inputs, no memory of past inputs. RNNs have the capacity to remember information about past to generate output which can be clear after unfolding the loop.

**LSTM :** LSTM (Long Short Term Memory) are an improvised form of the traditional RNNs. Recurrent Neural Networks suffer from 'Vanishing Gradient Problem', a problem where the network in unable to emanate the necessary gradient information from the rear end of the network to the layers near the front of the model. They are unable to handle long-term dependencies. On the bright side, LSTMs do not suffer these problem.

Introduced by Hochreiter & Schmidhuber [8] in 1997 A.D, LSTMs were tuned and famed so hence are equipped to handle long-term dependencies which the reason for me to choose the model for the research.

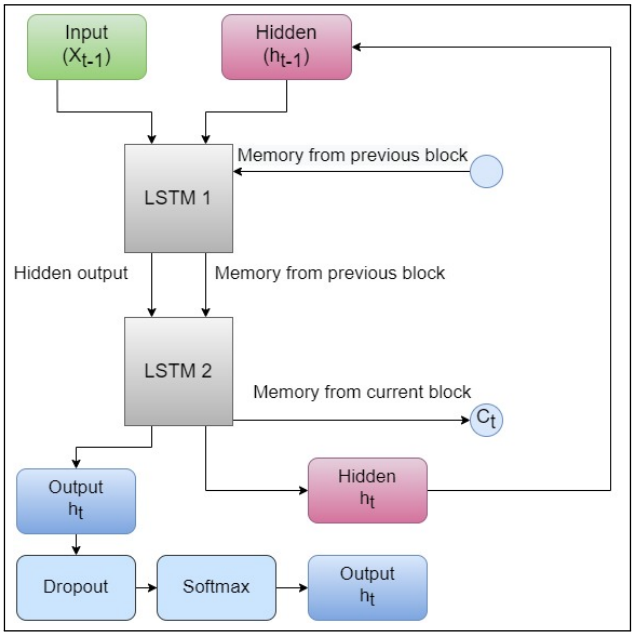


**Figure 2:** Architecture of the Network

Above given figure illustrates the architecture of the model. There are two layers of LSTM stacked together. I used batches of different sequence length as inputs into the model and trained the model. Input at a given time 't' depends on the current input as well as all the previous inputs.

### 4.4 Model Training and Music Generation

LSTM model is then trained on the string data which is obtained after the processing. With a sufficiently trained model, output sequence of string is predicted which is similar to the once that was fed during the training phase. All the input midi songs is converted to a sequence of format : *Notes-Duration-Next*.

*Notes* represents the note that is played, *Duration* represents for how long the note is played and *Next* represent the next note that will be played after t duration of time in milliseconds. Thus obtained sequence is then turned back to CSV file after processing and finally back to MIDI using a csvmidi program. The model produces desired length of sequence of characters which is then turned to MIDI after processing.

#### 4.4.1 Training on tracks from two composers.

After preprocessing of the tracks, the suitable CSV dataset was supplied to the model for training.Music from two composers from the classical period were

chosen for dataset collection: Ludwig van Beethoven, a german composer and Wolfgang Amadeus Mozart, an austrian composer. Both the composers are still considered legendary and are still prominent characters of recent world. Both the artist have provided absolute masterpieces, widening the horizon of sonata,concerto,quartet and symphony. After the training, a curve as seen in figure 3 is obtained.

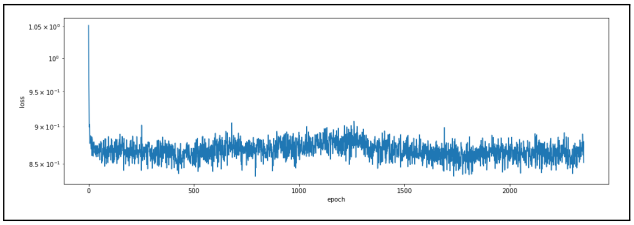


**Figure 3:** Loss vs Epoch curve for the Model on training up to 2353 epochs

we can deduce that losses from epoch range 1000 to 1500 seem to increase then any other interval. In general it is evident the losses keep fluctuating between 0.9 and 0.8 over all the range of the epoch. However a higher loss does not necessarily mean a bad model, it simply means the model is able to generate diverse notes from the training set.

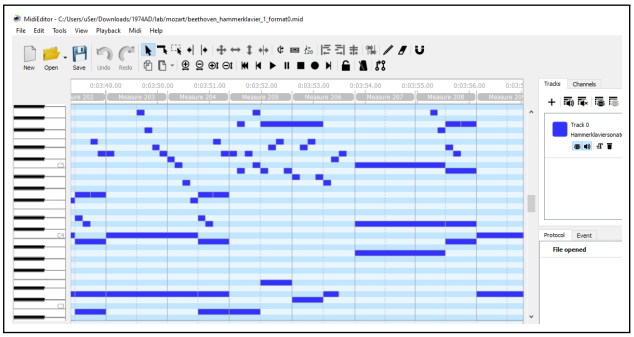

**Figure 4:** Midi Sequence of a random Bethoveen's song in the training set

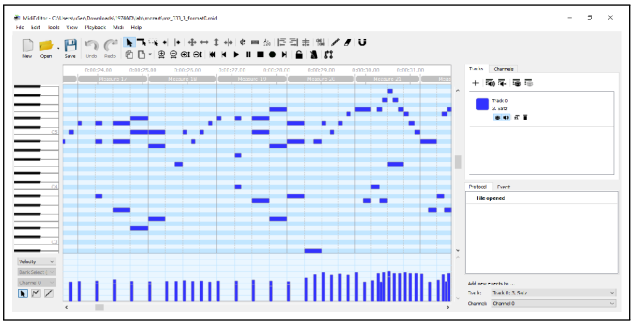

**Figure 5:** Midi sequence of a random Mozart's song in the training set

Upon looking at the MIDI sequences of the training tracks, variety of notes of different duration and interval is observed. Each composer has his own style of musical note creation. This difference in musical style is of key prominence as the model learns to produce a wide scope of notes in the song. Figure 6 is the midi sequence of one of the generated song by a trained model.

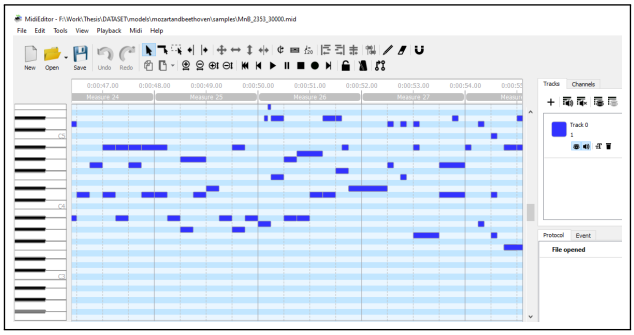

**Figure 6:** Midi sequence generated by the LSTM Model

On a close inspection, it can be seen that different notes of varied duration is generated by the model on its own. The model is able to generate sequences after learning from the input dataset.

### 4.4.2 Training on video games track

Video games track have piano music in them, Next step was for the training of the model on tracks collected from popular video games. After preprocesing, LSTM model was trained with the dataset and the training plot is shown in figure 7.

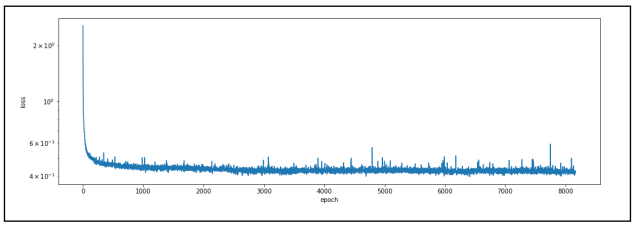


**Figure 7:** Loss vs Epoch graph on training LSTM model up to 8164 epoch

Loss seems to decrease from the first iteration to two thousand epochs and then the model does not seem to decrease the loss further. Video games music on general have a repetitive pattern of music with some variations now and then. These patterns can be vividly distinguished upon listening. Below given is an example of one of such pattern in the training set.

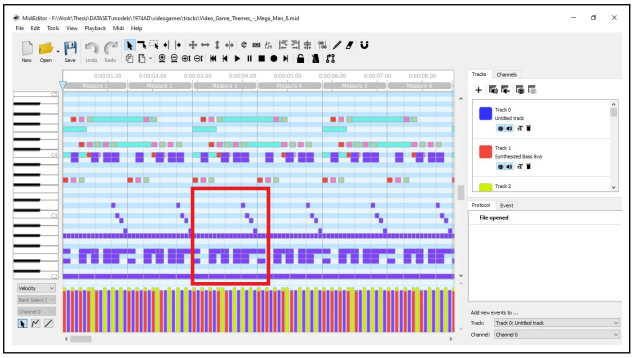

**Figure 8:** Repetitive patterns in the training set

A clear repetition pattern of music in one of the tracks as shown by a red rectangle is seen. Such patterns can often be seen in the video games track. A similar pattern is generated by the trained model as shown in figure 9.

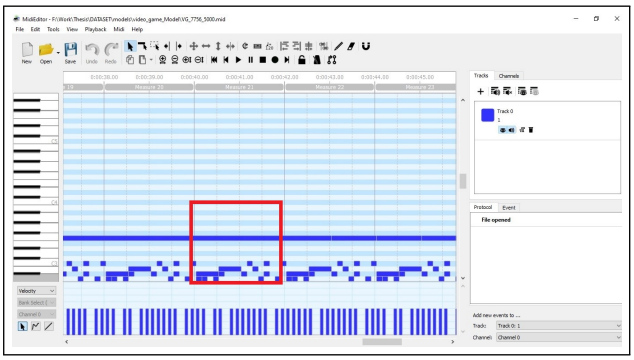

**Figure 9:** A repetitive pattern generated by the LSTM model

### 4.4.3 Plotting spectrogram

A spectrogram is a visual representation of the ”loudness” or signal strength over time at different frequencies contained in a specific waveform. Spectrogram for original musical piece and generated musical piece were plotted.

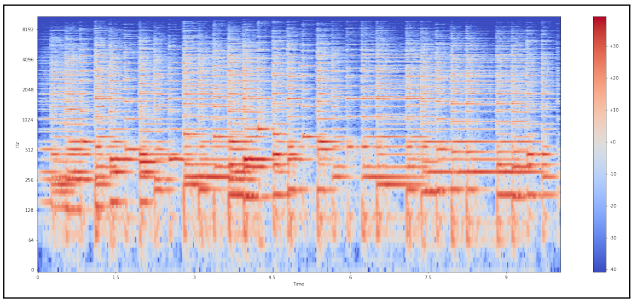

**Figure 10:** Spectrogram of short clip of a training track

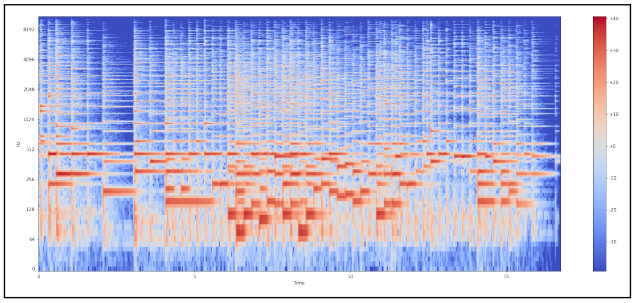

**Figure 11:** Spectrogram of short clip of a generated track

The frequency is represented by the vertical axis, with the lowest frequencies at the bottom and the highest frequencies at the top. The horizontal axis of time runs from left to right. The third dimension, color, represents the amplitude (or ”loudness” or energy) of a specific frequency at a specific time, with dark blues indicating low amplitudes and brighter colors up through red corresponding to progressively stronger (louder) amplitudes.

## 5. Music Evaluation

For the sake of music evaluation, an approach inspired from the paper [14] was chosen. As the paper had similar results comparing this research, the approach seemed logical and well founded. A total of six musical compositions sampled for about 15 to 20 seconds were taken, similar approach to the paper, out of which three compositions were automatically generated by one of the AI model and the other three were chosen from original Beethoven music.

An online evaluation form was created with the links to the compositions.A total of thirteen questions were asked to each respondent and their responses were recorded. A link to the music was present in the question from which the respondent could listen to the soundtracks. They could only listen to the tracks without any other information. Each musical composition was evaluated in terms of three evaluation indicators using a Likert scale. Similarly, the respondents were asked to choose the music that they listened to was composed by an AI or a human. Several musicians, music teachers, singers and the members of music society were requested to convey their responses.

Three evaluation indicators : Melody, Rhythm and Harmonic Interval were chosen and people were given a short introduction to those indicators and advised to

**Table 1:** Musical Composition Evaluation Score

| Music | Order | Author | Melody Score | Rhythm Score | Interval Score | Overall Score |
|---|---|---|---|---|---|---|
| M5 | 1 | Beethoven | 3.625 | 3.525 | 3.75 | 3.633 |
| M4 | 2 | LSTM | 3.675 | 3.55 | 3.575 | 3.6 |
| M3 | 3 | LSTM | 3.6 | 3.475 | 3.6 | 3.558 |
| M2 | 4 | Beethoven | 3.525 | 3.475 | 3.375 | 3.458 |
| M6 | 5 | LSTM | 3.325 | 3.325 | 3.525 | 3.392 |
| M1 | 6 | Beethoven | 3.325 | 3.3 | 3.175 | 3.267 |

evaluate the musical compositions as per their liking and understanding of the music.Each indicator had a total of five levels ranging from low to high: very poor, poor, average, good and excellent.

Online evaluators were only given the link to the music, not the composer's information. And the responses from the evaluators were recorded. Table 1 shows the scores of the six musical compositions.

Looking at the following results of musical evaluation, the compositions created by the LSTM model ranked second, third and fifth respectively. It can be inferred that the model can generate music that has a good melody, rhythm and harmonious interval.Similarly, a turing test to distinguish the musical compositions by Beethoven and the AI model is shown in figure 12.

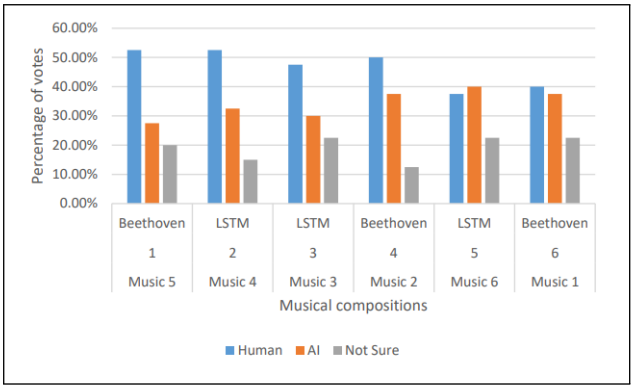


**Figure 12:** Turing Test Score

A visualization of the online musical evaluation responses shows that both the AI composed musical pieces Music 3 and Music 4 pass the test with scores of 47.5% and 52.5% for human composed criteria, respectively. However, Music 6 fails the test with a score of 40.0% for AI composed criteria. To back this up, 22.5% of people are unsure whether the composition was created by AI or by humans. It is evident from the following results that the composed music is human like as the participants are not clearly able to find the difference. Based on the responses collected, test result on whether to distinguish the musics composed by Human from that composed by an AI is shown in table 2.

**Table 2:** Distinguished as music composed by AI

| Music 1 | Music 2 | Music 3 | Music 4 | Music 5 | Music 6 |
|---|---|---|---|---|---|
| 37.50% | 37.50% | 30.00% | 32.50% | 27.50% | 40.00% |

Above results demonstrate that the model is able to generate music that can not be easily distinguished as machine composed.

## 6. Conclusion

This paper provides a novel way to create music based on deep learning. Key works of this paper are as follows:

1. Record and standardize music information in the form of note and duration sequences by gathering, analyzing, and extracting a lot of data about piano music.
2. Applying varied methods to preprocess the data and then via numerous trials based on LSTM and training, eventually a more suitable network model is obtained, and then automatic composition of music by the model.
3. Evaluation of the music through online survey indicates that the LSTM model can generate music with good musical qualities and It is challenging for average people to identify the true inventor of the music.

## 7. Future Works

Models currently play for a set amount of notes before abruptly stopping because they are unable to construct beginnings and endings to pieces. Non-abrupt endings might be possible if a method for the model to dynamically choose when to terminate a piece was developed.

Similarly, the model does not differentiate between notes and chords. A C major triad, for instance, is not thought of as a collection of notes, but rather as a single, distinct note altogether.

The network may get a deeper grasp of music if it was given the opportunity to learn how notes can combine to produce chords.In addition to notes, duration and offset, intensity and emotion are other key factors to consider. A deeper understanding of music theory and its application for further improvement of the model will be carried out.

## References


[1] Michael Conner, Lucas Gral, Kevin Adams, David Hunger, Reagan Strelow, and Alexander Neuwirth. Music generation using an lstm. *arXiv:2203.12105 [cs, eess]*, Mar. 2022.

[2] Iannis Xenakis. The origins of stochastic music. *Tempo*, pages 9–12, 1966.

[3] David Cope. *The Algorithmic Composer*. A-R Editions, Inc., Sept. 2000.

[4] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. Deep learning techniques for music generation – a survey. *arXiv:1709.01620 [cs]*, Aug. 2019.

[5] Mark J Steedman. A generative grammar for jazz chord sequences. *Music Perception: An Interdisciplinary Journal*, 2:52–77, Oct. 1984.

[6] Kemal Ebcioglu. An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12:43, 1988.

[7] Rebecca Fiebrink. Machine learning education for artists, musicians, and other creative practitioners. *ACM Transactions on Computing Education*, 19:1–32, 9 2019.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, Nov. 1997.

[9] Google. Magenta, 2020.

[10] Innovating for writers and artists – spotify for artists, May 2017.

[11] Serkan Sulun, Matthew E P Davies, and Paula Viana. Symbolic music generation conditioned on continuous-valued emotions. *IEEE Access*, 10:44617–44626, 2022.

[12] Clement Bertrand, Frederic Peschanski, Hanna Klaudel, and Matthieu Latapy. Pattern matching in link streams: Timed-automata with finite memory. *Scientific Annals of Computer Science*, 2018:161–198, Sept. 2018.

[13] Mnist handwritten digit database, yann lecun, corinna cortes and chris burges, 2009.

[14] Minghe Kong and Lican Huang. Bach style music authoring system based on deep learning. *arXiv:2110.02640 [cs]*, Oct. 2021.