Symbol-Based End-to-End Raw Audio Music Generation

Changxuan Wu, Ting Lan, Chunyan Yu, Xiu Wang*

College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
*wangx@fzu.edu.cn

Abstract. In recent years, deep learning has emerged in the audio field with many excellent models and beats non-depth methods in the quality of generated audio. This paper implements a symbol-based end-to-end music generation model. This model generates piano music corresponding to the pitch of the musical score using a two-dimensional “Piano-roll” liked structure as input. The experiments show the generated music obtains good performance and achieves a result similar to the original song in pitch, melody, and timbre. Compared with other generation methods, the input of our model is simple, easy to obtain, and can generate music through an end-to-end method.

1. Introduction

The current music generation tasks are mainly divided into waveform-based music generation and symbol-based music generation [3]. For waveform-based music generation task, because the waveform details of the music are too rich and the representation space is continuous, it is difficult and unstable to extract the features contained in the waveform from the music. The music generated by this method containing more noise. The symbol-based music generation task uses symbols to represent musical score information, which reduces the difficulty of the model in understanding the musical information and makes it easier to extract musical features. For the symbol-based generation model, the input can be simple, and the input is generally a character representation of a musical score, that is, a one-dimensional or two-dimensional character sequence. And there is no requirement for the length of the character sequence, and it can be any length. This method has a relatively simple model structure and can achieve end-to-end music generation without the need to train multiple models separately.

In recent years, there have some advanced music generation models that use a symbolic representation of musical score information as input and get music audio waveforms. Hongyuan Zhu et al.[16] generate music by combining symbolic representations of melody, chord, and rhythm information based on knowledge of music theory. [16] points out that using the GRU network to combine this basic music theory information can get good performance. The GRU network is more suitable for music generation tasks than other network structures. Because it has a recurrent neural network structure that can perceive context information well, which is the key point of the music generation task. However, the complex input structure is not conducive to music generation and its derivative tasks, such as automatic composing tasks, and it is difficult to obtain aligned music theory information such as chords from the score. We hope to generate corresponding music through a simple musical score representation.

Symbol-based music generation (Score to Music, STM) is the conversion of note sequences to music, and Speech synthesis (Text to Speech, TTS) is the conversion of text sequences to speech, so...
they are very similar in structure. Therefore, we can develop symbol-based music generation methods based on the method of speech synthesis. In fact, many scholars have been doing such work [3][5][6][15]. Tacotron is an end-to-end speech synthesis model proposed by Yuxuan Wang et al.[14] in 2017. The model takes the character sequence as input, gets the corresponding speech spectrum map, and then predicts the phase through the Griffin-Lim algorithm to obtain high-quality speech audio. The model does not use the depth model as the vocoder but chooses the Griffin-Lim algorithm, which makes the model’s speech generation more controllable. The model generation task is similar to the task of this paper in the framework. It also uses simple input to generate sound fragments corresponding to the input. Therefore, we use Tacotron’s model framework to adapt it to music generation tasks.

2. Method

2.1. Architecture
Most music generation models are roughly composed of three parts: encoder, decoder, and vocoder. Our model design also follows this structure. A post-processing network was applied between decoder and vocoder, because the linear-spectrogram used by vocoder contained rich details, so we use mel-spectrogram as a transition. Shown in Fig.1 was our structure.

We using the CBHG structure as the model’s encoder, because CBHG can extract abstract high-dimensional features from the context of the sequence well. [14] pointed out that the CBHG structure can not only reduce overfitting in speech synthesis tasks but also reduce wrong pronunciation compared with the traditional multilayer recurrent neural network (RNN) structure. CBHG consists of a bank of 1-D convolutional filters, a highway networks, and a bidirectional gated recurrent unit (GRU).

We divide the decode operation into two steps, firstly generate Mel-spectrogram from the music score features, then convert Mel-spectrogram into linear-spectrogram with post-processing net. Because linear-spectrogram is too complicated with a lot of information, it is more difficult to generate linear-spectrogram directly through the decoder. Since the lengths of the input score features and the spectrum features are not the same, the decoder mainly uses a dynamic double-layer GRU to generate the spectrum features of the corresponding length. At the same time, the local sensitive attention mechanism mentioned in [1] makes the recurrent neural network’s perception range of input increase, because this attention mechanism can combine the focus position of the previous step and the characteristics of the input sequence [10]. When training the dynamic double-layer GRU, we randomly using target Mel-spectrogram as input frame to help training, which is also shown in Fig.1. Whether to use the target spectrogram is decided by a threshold which is decreases along time.

![Fig.1 Our model structure](image)

We use the modified CBHG as a post-processing network to obtain a linear spectrum. The Mel-spectrogram is obtained from a linear spectrogram using a set of mel filters and has less detail
information. However, the two have the same time dimension, but the richness of the frequency information contained in each moment is different. We use CBHG to extract high-dimensional features to obtain a linear spectrogram. Since the difference between the two spectrums in the time dimension is small, we found that the bidirectional GRU in the CBHG structure has little improvement in the quality of the spectrum, and increases the model training time. Therefore, we remove the bidirectional GRU structure, but retain the Highway network structure, because it can increase the depth of the network and the model will not become untrainable.

This model uses the Griffin-lim algorithm to generate music waveforms, because the experiments in [8] show that there is no enough gap between the deep network, such as wavenet and waveglow vocoder, and the Griffin-lim algorithm in the voice quality. And Sheng et al.[11] mentioned that the closer the spectrogram generated by the model to the original spectrogram, the better the audio quality, whether it is to use the vocoder of the deep model or the algorithm. So we are more focused on improving the quality of spectrograms.

2.2. Representation of music score
In the research of symbol-based music generation, the input data used, that is, the musical notation information represented by symbols, are different. [16] used chord, rhythm, and melody information as input to the model. Such input is more complicated and it is difficult to get accurate input information. Fig.2 shows the result of visualizing this structure as a picture. The horizontal axis is the time axis, and the vertical axis is the n-level pitch expression which can basically map the pitch of the notes in the score. In such a two-dimensional structure, each point represents the pitch of the note corresponding to the current moment and the velocity of the note. Such a structure is a good representation of polyphony, and it can be expressed accurately no matter how many keys the pianist presses at the same time. We regard it as a $t \times n$ matrix, where $t$ represents time and $n$ represents pitch, typically, $n=128$.

![Fig.2 “Piano roll”. This is a piano roll you would typically see in a digital audio workstation [9].](image)

3. Experiments and evaluation
Assessing the quality of artificially generated music is a very challenging task. Even the concept of quality is very difficult to define uniformly, because it is completely subjective, and everyone has their own standards and definitions for music quality evaluation. We mainly use subjective evaluation indicators to evaluate the audio quality, naturalness, pitch, and melody generated by the system1.

3.1. Dataset
Our experiment uses Maestro2 [4] dataset. The dataset contains over 200 hours of paired audio and MIDI recordings from ten years of International Piano-e-Competition. Audio and MIDI files are aligned with about 3-millisecond accuracy and sliced to individual musical pieces, which are annotated with composer, title, and year of performance.

3.2. Setup
We preprocess the dataset so that the duration of each piece of music is about 5-second. For the input piano roll, we use a sample rate of 100 frames per second (fps) to calculate it using the corresponding

---

1 All the music sample can be found in: https://aberaven.github.io/
midi file and the pitch in the piano roll is absolute pitch. The output sampling rate of the model is set to 44100 Hz. The generative music used in all experiments was generated from models trained with 80k steps.

3.2.1. Chromagram Comparation
Chroma feature is a quality of tone level, referring to the “color” of musical tones, which can be broken down into an octave constant (called “chroma”) and a “pitch” height, and this height indicates the octave where the pitch is [12][13]. Hence, the chroma feature, which is a sequence of chroma vectors, can roughly see the similarity of the pitch of the two pieces of music. We set a threshold on the result to remove the interference components with smaller values.

3.2.2. ABX Test
This method is an improvement on the pair comparison method. Testers are required to listen to audio signal A (which is the original song), then audio signal B (which is randomly selected), and finally audio signal X in a sequence of time. The tester’s task is to determine whether audio signal X is more like audio signal A or audio signal B.

For this study, we randomly select 20 piano music fragments from the test set, with a length of 5-second, and generated music through the music score corresponding to these fragments. We define the similarity between two pieces of music as the similarity in melody, rhythm, pitch, and timbre. The higher the accuracy, the closer the model generated music is to the original music.

3.2.3. MOS Test
The mean opinion score is a common measurement used in the evaluation of video, audio, and audiovisual quality, and represents the overall quality of the system. This experiment using three evaluation experiments: whether A and B have the same melody, whether A and B have the same pitch, and evaluate the quality of the generated music clip.

All experiments using the same 20 5-second music generated-original pair. MOS is expressed as a single rational number, usually in the range of 1-5 (1-Strongly disagree/Bad, 2-Disagree/Poor, 3-Neither agree nor disagree/Fair, 4-Agree/Good, 5-Strongly agree/Excellent). The purpose of the above three experiments is to show whether the model can generate the same melody and pitch as the real music audio and to evaluate the quality of the music generated by the model.

3.3. Results and analysis

3.3.1. Is it possible to generate music with the same pitch as the source music?
Then, we calculate the similarity of the histogram of the source music chromagram and the model generated music chromagram to evaluate the pitch accuracy of the model generated music and the source music. We obtain an average histogram similarity of 300 generated music fragments and source music fragments of 0.8367. Compared with the source clip, the generated music clip also has a lower MSE, and the average result of 300 test audios is 0.11. It can also be seen from Fig.3 that the chromagram of the original song and the chromagram of the generated music are relatively close, that is, the pitch level is matched. So the generated music has a similar pitch level as the source music. Not only that, but it can also be seen from the pitch MOS test, shown in Table.2, that the model has achieved good results in subjective evaluation experiments. This shows that the music generated by the model has high consistency in terms of human subjective perception with the pitch of the source music.

3.3.2. Can the model generate music with the same melody as the source music?
The results of the ABX test experiment are shown in Table.1. From the table, we can see that more than 95% of the testers believe that the generated music is similar to the source music, which proves
that the model can generate music similar to the original music, and the pitch of the generated music is almost the same as the pitch information in the score representation of the model input. However, we

found that the model achieved poor performance on test music 7. It was found through listening to the music that the source song included not only piano music but also the sound of the violin. And the sound of the violin is louder than that of the piano, which results in the generated music not sounding like the source music. But the music produced by the model has a similar melody to the piano music in the background of the original music. It can be seen from the experimental results in Table.2 that in the subjective evaluation, the music generated by the model achieved good results. The testers acknowledge that the music generated by the model has the same melody as the source music. It can be seen that the music generated by the model has the same melody as the music corresponding to the score.

Table.1 ABX Test result

| Test music | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| A          | 5 | 5 | 4 | 5 | 5 | 5 | 2 | 5 | 5 | 5  | 5  | 5  | 5  | 5  | 5  | 5  | 5  | 5  | 5  | 5  |
| B          | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

3.3.3. Is the generated music of high quality?

It can be seen from the experimental results in Table.2 that the model also achieves good results in terms of the quality of generated music. However, for some music with a higher pitch, the generated music is partially distorted and does not sound like piano music. Because the tone of the instrument is determined by the harmonics, then the model fits smoothly to the harmonics, and is difficult to fit the high-frequency harmonics, so the high-frequency piano music does not sound like a natural piano to the human ear and artificial traces in that music are more serious.

Table.2 MOS scoring result

| Experiment | Melody MOS | Pitch MOS | Quality MOS |
|------------|------------|-----------|-------------|
| Score      | 4          | 3.7       | 3.5         |
4. Conclusion

In this paper, we propose an end-to-end single-instrument music generation model based on symbols. It can be seen from experiments that our method can generate smooth and high-quality music fragments. The generated music fragments have similar melody and pitch to the original music, and there is no burr or noise in quality. The quality of the generated music is high and it can generate music with corresponding pitch, but the simulation of the model for harmonics is still different from the actual harmonics. The input of the model proposed in this paper is simple and provides more possibilities for music generation and its extended task.

Acknowledgments

This work was supported by the Fujian Foundation of Science 2018J01794 and the Fujian Health-Education Joint Project WKJ2016-2-26.

References

[1] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in neural information processing systems. pp. 577–585 (2015)
[2] Commission, I.E., et al.: Electropedia: The world’s online electrotechnical vocabulary. electropedia.org (2010)
[3] Eppe, M., Alpay, T., Wermter, S.: Towards end-to-end raw audio music synthesis. In: International Conference on Artificial Neural Networks. pp. 137–146. Springer (2018)
[4] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the maestro dataset. arXiv preprint arXiv:1810.12247 (2018)
[5] Kim, J., Choi, H., Park, J., Hahn, M., Kim, S.J., Kim, J.J.: Korean singing voice synthesis based on an lstm recurrent neural network. In: Interspeech. pp. 1551–1555 (2018)
[6] Manzelli, R., Thakkar, V., Siahkamari, A., Kulis, B.: Conditioning deep generative raw audio models for structured automatic music. arXiv preprint arXiv:1806.09905 (2018)
[7] McLeod, P., Wyvill, G.: A smarter way to find pitch. In: ICMC. vol. 5, pp. 138–141 (2005)
[8] Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flow-based generative network for speech synthesis. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3617–3621. IEEE (2019)
[9] Romo, T.: Midi: A standard for music in the ever changing digital age (2018)
[10] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783. IEEE (2018)
[11] Sheng, L., Huang, D.Y., Pavlovskiy, E.N.: High-quality speech synthesis using super-resolution mel-spectrogram. arXiv preprint arXiv:1912.01167 (2019)
[12] Shepard, R.N.: Circularity in judgments of relative pitch. The Journal of the Acoustical Society of America 36(12), 2346–2353 (1964)
[13] Sumi, K., Arai, M., Fujishima, T., Hashimoto, S.: A music retrieval system using chroma and pitch features based on conditional random fields. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1997–2000. IEEE (2012)
[14] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al.: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
[15] Yang, L.C., Chou, S.Y., Yang, Y.H.: Midinet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847 (2017)
[16] Zhu, H., Liu, Q., Yuan, N.J., Qin, C., Li, J., Zhang, K., Zhou, G., Wei, F., Xu, Y., Chen, E.: Xiaoice band: A melody and arrangement generation framework for pop music.
Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2837–2846. ACM (2018)