Music generation and human voice conversion based on LSTM

Guangwei Li[1], *Shuxue* Ding[1,*], *Yujie* Li[1], and *Kangkang* Zhang[1]

[1]School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China

Abstract. Music is closely related to human life and is an important way for people to express their feelings in life. Deep neural networks have played a significant role in the field of music processing. There are many different neural network models to implement deep learning for audio processing. For general neural networks, there are problems such as complex operation and slow computing speed. In this paper, we introduce Long Short-Term Memory (LSTM), which is a circulating neural network, to realize end-to-end training. The network structure is simple and can generate better audio sequences after the training model. After music generation, human voice conversion is important for music understanding and inserting lyrics to pure music. We propose the audio segmentation technology for segmenting the fixed length of the human voice. Different notes are classified through piano music without considering the scale and are correlated with the different human voices we get. Finally, through the transformation, we can express the generated piano music through the output of the human voice. Experimental results demonstrate that the proposed scheme can successfully obtain a human voice from pure piano Music generated by LSTM.

1 Introduction

Music, as an art form expressing emotions, is in high demand in the market. At present, the number of professional music creators is limited, and music production takes time and effort, and the cost is high[1]. In recent years, with the rapid development of artificial intelligence deep learning algorithms in image recognition[2], video detection[3], natural language processing[4], and speech processing[5] and its application in various fields, computer music has brought us a new opportunity for development. The development of deep learning in music allows us to use models to generate piano music. We use a deep learning based algorithm to compose and generate music on the computer.

Music generation is a study that uses algorithms to automate part or all of the music creation process. Although the study of music generation using mathematical methods has been around for a long time, it has not made great progress due to the limitation of the development of other related disciplines. In recent years, with the development of deep learning, the music generation problem has come back into our field of vision. Various related models have also been applied to the study of the music generation problem.

* Corresponding author: sding@guet.edu.cn

Recurrent Neural Network (RNN) is an efficient model designed to process sequential or temporal data[6]. However, due to the problems of long-term dependence and gradient disappearance, we adopt an LSTM network to design the model, which can not only process the information on multiple scales but also process fine-grained temporal details, as well as coarse-grained remote historical information. LSTM adds the idea of self-circulation to keep the gradient flowing, which can effectively solve the problems of long-term dependence and gradient explosion[7].Through the use of the LSTM music generation model to train a large number of piano music, automatic generation of new piano music. The forecast distribution of this model is not only related to the current state but also related to the previous state to some extent. Through LSTM, we can get more fluent piano music, which is better than other generated models.

This paper mainly uses a mature model of piano music generation—LSTM and then converse piano music to the human voice[8][9]. We have collected a large number of piano MIDI data sets and obtained piano pieces of any length through LSTM training. Model training based on MIDI data sets is to predict the next note in a sequence based on the current combination of note and chord information. Because our experiment is currently only able to perform voice conversion for single-key piano repertoire. We perform a secondary screening of piano music to extract the piano music with the single key-value we need. Through the corresponding position relationship with our human voice segment conversion. The transformation between piano music and our voice information lays a foundation for the next step to realize the automatic generation of music with complete lyrics and emotions.

2 Related works

2.1 Music expression

The choice of an encoding form of music information is closely related to the processing of input and output of the depth frame. The data forms of music mainly include audio and symbolic, which correspond to the division of continuous variables and discrete variables respectively.

Music information based on audio is mainly represented by a signal wave and spectrum. This kind of music representation can retain complete original music information, but it has some shortcomings, such as the large consumption of computing resources and long processing time. For another kind of music information based on symbols, the data is mainly converted into the form of symbols, the representative of which is piano music. During the data preprocessing of piano music, the data is converted into the form of a one-hot vector for data processing. According to the key position of the piano and the time step of playing, we converted music data into the form of a one-hot vector to serve as the training data of our model training.

2.2 Existing methods for music generation

Applications in music generation have been studied for a very long time. Back in the 1980s, Steedman used a small number of rules to generate a large number of complex chord sequences[10]. Pachet proposed a method to control the Markov model for melody generation by a specific method[11].

Karen Simonyan and Sander Dieleman generated a piece of piano music audio through processing the audio waveform through the WaveNet Deep learning network[12]. For WaveNet, the model is fully autoregressive and the predicted distribution of each audio sample depends on all previous samples. WaveNet combines causal filters with extended

convolution, which enlarges their receptive field but still creates a certain amount of noise in the generated audio. To improve the reliability of the network and the quality of the obtained audio, an RNN method is proposed. RNN has a better effect on processing our data set than the general neural network, but it can only process short-term data because of the gradient disappearance problem.

To solve those problems existing in the recurrent network, Eck et al. use LSTM to generate music, their experimental results show that LSTM learns a form of blues music successfully and can compose novel melodies in that style[13].

3 Model and formulation

In this paper, we combine the LSTM generation model with the audio cutting technique. The obtained piano music sequence is screened and converted to the human voice to obtain a piece of audio information of the human voice.

For the whole process of obtaining human voice information from the piano music, we obtained the experimental results we wanted through the above techniques. The overall flow chart of piano music and human voice conversion is shown in Fig. 1.

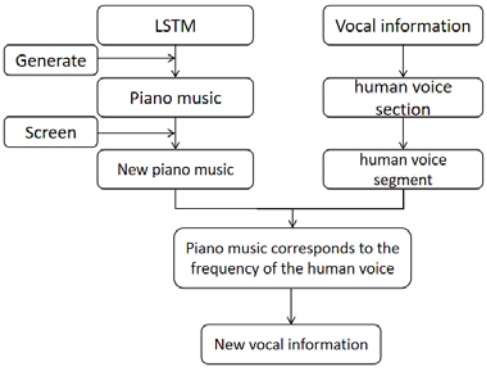

Fig. 1. Overall flow chart of piano music and human voice conversion.

We collect voice information for different notes according to the different sounds made by piano music. We collect the corresponding human voice audio according to the different notes. Then, the whole human voice audio is cut through the audio segmentation technology to each audio segment.

For piano music generation, we use the LSTM network model to generate the piano music we need. Then, we use the piano music generated from LSTM for human voice transformation.

3.1 Human voice section

For the human voice transformation, we collected the human voice information for piano music and human voice transformation. Ignoring the pitch of the piano keys, the main melody of piano music can divide into seven different notes, which are regarded as seven different pieces of audio from 1 to 7. These 7 sounds are collected as our human voice audio information. We propose two solutions for the human voice information we need.

In the first scheme, the voice information corresponding to each note is separately collected (Individual recording, IR), as shown in Fig. 2. We can directly get the audio information of each segment without processing the audio file through the audio slicing technology.

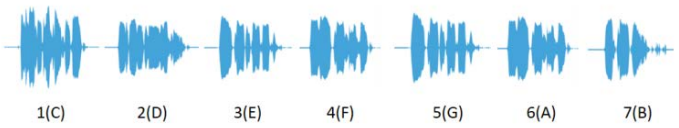

Fig. 2. Individual recording (IR): The audio file is collected separately.

In the second scheme, we collect the human voice information into an audio segment (Overall the recording, OR) to ensure the fluency of the audio segment. Then, audio slicing technology is applied to the collected audio files, as shown in Fig. 3.

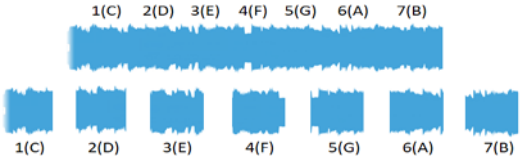

Fig. 3. Overall the recording (OR): Audio files are collected and sliced.

3.2 Piano music generated from LSTM

The LSTM model has a very good effect on the generation of piano music sequences. We use the LSTM model to generate piano music to realize the mutual transformation between our piano music and the human voice, and its structure is shown in Fig. 4.

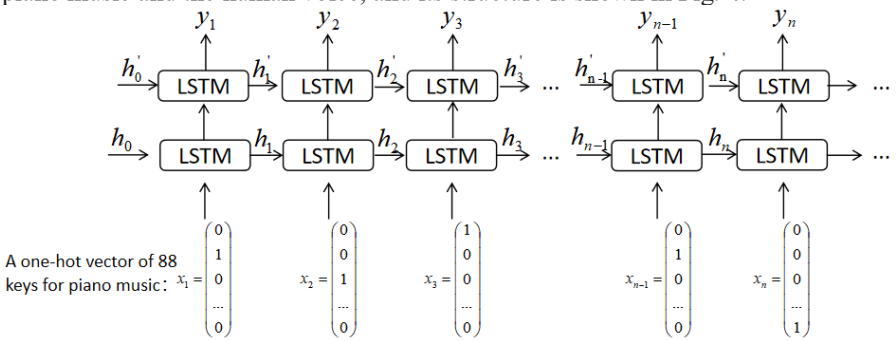

Fig. 4. LSTM structure for music generation.

There are four main gates in the LSTM model as follows,

- g_t: Input gate, Learns whether to write to cell.
- f_t: Forget gate, Learns whether to erase cell state.
- q_t: Output gate, Learns how much to reveal to the cell state.
- $\tilde{s}_t$: Update gate, Learns how much to write to cell state.

According to the following formulations, we can obtain Cell State s_t and Output State h_t.

$$\tilde{s}_t = \tanh(b + Ux_t + Wh_{t-1}) \tag{1}$$

$$q_t = \sigma(b_q + U_q x_t + W_q h_{t-1}) \tag{2}$$

$$f_t = \sigma(b_f + U_f x_t + W_f h_{t-1}) \tag{3}$$

$$g_t = \sigma(b_g + U_g x_t + w_g h_{t-1}) \tag{4}$$

The equations above describe a single layer of LSTM. We can have a stacked LSTM structure where the output of layer-1 acts like the input of layer-2 and so on[14].

$$s_t = f_t * s_{t-1} + g_t * \tilde{s}_t \tag{5}$$

$$h_t = q_t * \tanh(s_t) \tag{6}$$

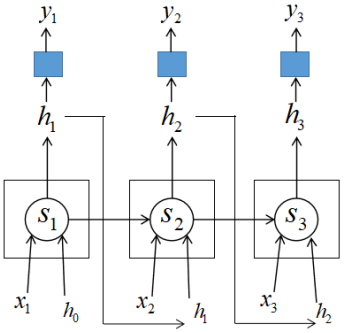

Fig. 5. The unrolled network of LSTM.

The unrolled network during the forward pass is shown in Fig. 5. Note that the gates have not been shown for simplicity. s_T is responsible for computing h_T as well as the next cell state s_{T+1}. At each time step, x_t and h_{t-1} obtain s_t and h_t according to the formula, and calculate y_t through the activation function.

4 Experiments

In the experiments, we trained and tested the proposed model. We generate piano music from the LSTM. The audio frequency of the human voice is obtained through the proposed IR or OR.

For piano music generation, we have collected a large number of piano music MIDI data files. By training the data set with the built model, we batch our sequence data during training, thus greatly speed up the training stage. However, as our sequence gets longer, we use more and more memory, so we need to solve the memory consumption problem[15]. In this paper, the gradient checkpoint method is used. This allows us to train the entire sequence of the model with less memory and perform more computations.

4.1 Training

Training piano music to generate random piano music sequences. Firstly, the piano music sequence is screened by simple single key data. Secondly, the filtered data is sliced and corresponds to the human voice audio frequency. Finally, the corresponding voice conversion is obtained. The corresponding relationship between our piano music and the sliced human voice audio is shown in Fig. 6. Through the position of each sound in the real

piano keys and the position in each column of the piano music matrix we obtained, we corresponded and numbered them sequentially. Through the key value of each tone, we get the human voice audio when we press the appropriate piano keys.

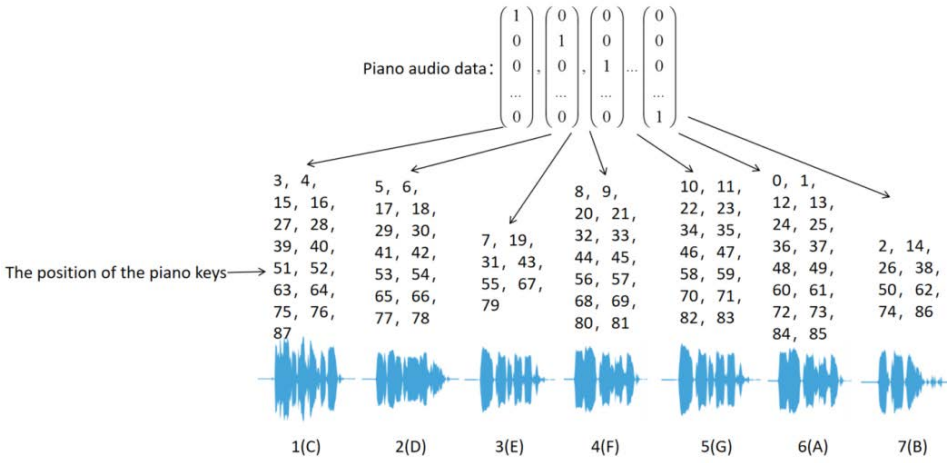

Fig. 6. The correspondence between piano music and the human voice.

After we have defined the corresponding relationship, the human voice information of piano music is obtained through the piano music sequence generated by the LSTM model. Then according to the piano music, we connected the corresponding audio segments into complete human voice information.

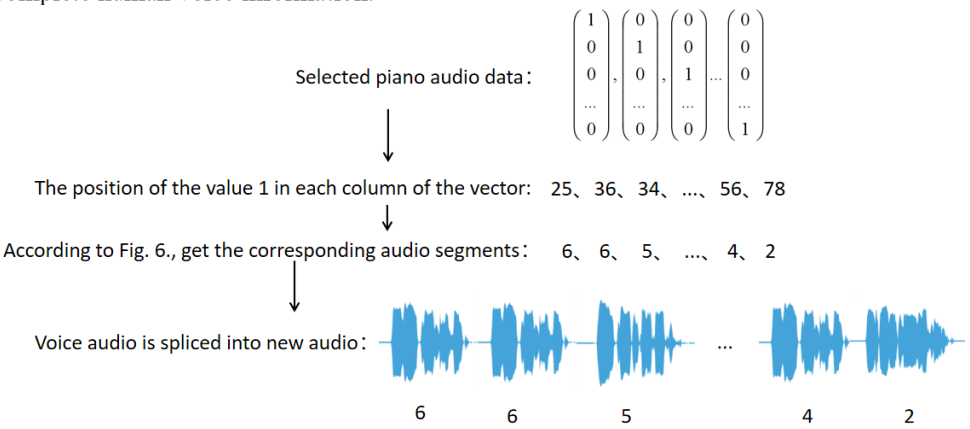

Fig. 7. The illustration of human voice conversion from piano music generated by LSTM.

According to Fig. 7, we generate piano music sequences based on the LSTM model and screened them. The selected piano music is corresponding through the relation specified by us, and audio blocks are obtained according to the position of key values. Finally, we spliced all the audio pieces according to the piano music to get the piano music vocal audio we wanted.

4.2 Experimental discussion

For our audio slicing technology, we propose two different ways of processing human voice information, IR and OR. By segmenting and slicing, we get the corresponding vocal

fragments of each note and then match the human audio fragments. The results obtained by the two different methods are shown in Fig. 8 and Fig. 9.

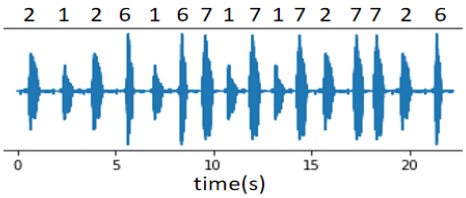

Fig. 8. The human voice audio waveform generated by IR.

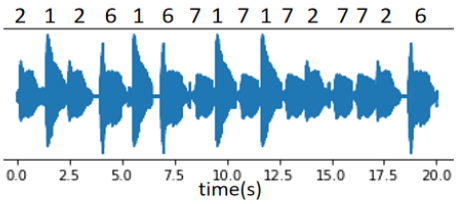

Fig. 9. The human voice audio waveform generated by OR.

Fig. 8 shows the experimental result of IR, although the integrity of each audio is guaranteed, there will not be two notes in the same audio. This method loses fluency after the audio splicing. Fig. 9 shows the experimental result of OR, from which we can see the voice audio waveform generated by OR is smoother. This method has strict requirements on the recording time. It requires each piece of audio after the slice contains only one note of voice audio. We can obtain a relatively good human voice audio through the method of OR.

5 Conclusions

In this paper, we proposed a novel approach for piano music and human voice conversion. we conducted a series of processing on the piano music sequence obtained by LSTM to obtain new data. Then, we matched the piano music sequence generated by the mature LSTM model with the human voice information we collected. In this way, we convert piano music to human audio and use human voices to represent piano music. In the future research, we plan to generate music with lyrics and emotions. Through the simple piano music and the conversion of the human voice proposed by this paper, the expression forms of human voice and melody can be understood and laid a foundation for the follow-up research.

References

1. Dannenberg RB. Music Representation Issues, Techniques, and Systems. *Computer Music J*, **17(3)**, 20-30 (1993).
2. A. Krizhevsky, K. Sutskever, G. Hinton. Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 1097-1105 (2012).
3. J. Ba, V. Mnih, K. Kavukcuoglu. Multiple object recognition with visual attention. *International Conference on Learning Representations* (2014).
4. T. Mikolov, A. Deoras, D. Povey, L. Burget, J. Cernocky. Strategies for training large scale neural network language models. *Automatic Speech Recognition and Understanding* 196–201 (2011).

5. T. Sainath, B. Mohamed, B. Kingsbury, B. Ramabhadran. Deep convolutional neural networks for LVCSR. *Acoustics, Speech and Signal Processing* 8614-8618 (2013).
6. H. Chu, R. Urtasun, S. Fidler. Song From PI: Amusically Plausible Network For Pop Music Generation. *Under review as a conference paper at ICLR 2017*. (2016).
7. S. Agarwal, V. Saxena, V. Singal, S. Aggarwal. LSTM based Music Generation with Dataset Preprocessing and Reconstruction Techniques, *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 455-462 (2018).
8. K. Zhao, S. Li, J. Cai, H. Wang, J. Wang. An Emotional Symbolic Music Generation System based on LSTM Networks, *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 2039-2043 (2019).
9. T. Jiang, Q. Xiao. Music Generation Using Bidirectional Recurrent Network, *2019 IEEE 2nd International Conference on Electronics Technology (ICET)*, pp. 564-569, (2019).
10. M. Steedman. A generative grammar for Jazz chord sequences. *Music Perception*, **2(1)**:52–77 (1984).
11. F. Pachet, P. Roy, G. Barbieri. Finite-length markov processes with constraints. *In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 635–642 (2011).
12. S. Dieleman, K. Simonyan. The challenge of realistic music generation: modeling raw audio at scale, *Computer Science,* (2018). URL https://arxiv.org/abs/1806.10474
13. D. Eck, J. Schmidhuber. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103 (2012).
14. F. Shah, T. Naik, N. Vyas. LSTM based Music Generation, *2019 International Conference on Machine Learning and Data Engineering (iCMLDE),* (2019).
15. J. Wang, X. Wang, J. Cai. Jazz Music Generation Based on Grammar and LSTM, *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 115-120 (2019).