An Emotional Symbolic Music Generation System based on LSTM Networks

Kun Zhao, Siqi Li
Information and Communication Engineering School
Communication University of China
Beijing, China
email: {zhaokun_213, lisiqi}@cuc.edu.cn

Juanjuan Cai, Hui Wang, Jingling Wang*
Key Laboratory of Media Audio & Video, Ministry of Education
Communication University of China
Beijing, China
email: {caijuanjuan, hwang, wjl}@cuc.edu.cn

***Abstract*—With the development of AI technology in recent years, Neural Networks have been used in the task of algorithmic music composition and have achieved desirable results. Music is highly associated with human emotion, however, there are few attempts of intelligent music composition in the scene of expressing different emotions. In this work, Biaxial LSTM networks have been used to generate polyphonic music, and the thought of LookBack is also introduced into the architecture to improve the long-term structure. Above all, we design a novel system for emotional music generation with a manner of steerable parameters for 4 basic emotions divided by Russell's 2-demonsion valence-arousal (VA) emotional space. The evaluation indices of generated music by this model is closer to real music, and via human listening test, it shows that the different affects expressed by the generated emotional samples can be distinguished correctly in majority.**

Keywords—algorithmic composition, music generation, emotional music, neural networks, LSTM

I. INTRODUCTION

In the computational creativity frontier, machine learning algorithms are more and more present in music creative domains and bring a brand-new way to help compose pieces or give inspirations for musicians. As German philosopher Hagel has said: "Music is the art of mood. It is straightly directed against mood." It is thus clear that connection is interwoven between music and human emotion. A model which can create music conditioned on emotion will bring lots of benefits to the field of film and composition. For instance, it could help movie makers with background music of specific emotion for different scenes so as to arouse resonance with audiences' feelings. And inspired by the music generated from the model, composers can make music for specific creative tasks, etc. Therefore, it is very meaningful to combine algorithmic composition technology with emotion. However, few attempts of music generation of different emotions have been found yet.

Music generation includes symbolic-domain generation (i.e. generating MIDIs, piano rolls) and audio-domain generation (i.e. generating Wavs). In this work, we only concentrate on generating MIDIs. Most existing neural network models for symbolic music generation use recurrent neural networks (RNN) or its variants, as they perform well in modeling temporal structure in music. Early in 1989, Todd used RNN to generate monophonic melodies which became the first attempt to apply artificial neural networks to music composition. Nevertheless, there is a weakness of "forgetting" distal events for RNN [1]. The Long-Short Term Memory (LSTM) networks [2], a variant of RNN proposed by Hochreiter and Schmidhuber has improved on this. And LSTM was first applied by Eck et al. in music composition so as to generate Blues monophonic melodies constrained on chord in 2002[3]. Since then, Music composition algorithm employing LSTM units is widely used [4,5,6].

Generally, polyphonic music is composed of both melody and accompaniment. And according to the composition skills, the accompaniment can be divided into two types which are chord and counterpoint. Therefore, polyphonic music has complex patterns along multiple axes: there are both sequential patterns between timesteps and harmonic intervals between simultaneous notes. Thus, Polyphonic music generation is more complicated than monophonic music generation.

Most prior works chose to simplify polyphonic music generation in certain ways to render the problem manageable. Such as generating only single-track monophonic music and then accompany it with chords conforming to composition rules. There are also some efforts to create harmonious chords for monophonic melodies. For example, Hyungui Lim et al. used bidirectional LSTM to generate chord sequence based on melody [7]. In Song from PI, Chu et al. employed a hierarchy of recurrent layers to generate not only melody but also drums and chords for the melody, leading to a multi-track pop song [8]. In addition to the methods above, there is another way which model polyphonic music as joint probability distributions of the combination of notes. The model can be trained to learn the musical pattern from collections of existing music, and then generate polyphonic music with both melody and accompaniment at the meantime. In the early stage, RNN-RBM, which proposed by Boulanger-Lewandowski, was able to generate polyphonic piano-rolls of a single track in 2012 [9]. Based on above architecture, a convolutional RBM with constraints was introduced for imposing higher-level structure on generated polyphonic music [10]. DeepBach proposed by Sony CSL, was specifically designed for composing polyphonic four-part chorale music in the style of Bach [11]. Some latest works have also begun to explore using GANs to create music [12,13,17]. Moreover, A network architecture

978-1-5386-6243-4/19/$31.00 ©2019 IEEE

proposed by Daniel D. Johnson, Bi-axial LSTM (BALSTM), is capable of modeling the joint distribution of notes while maintaining transposition invariance [14] Inspired by Bi-axial LSTM, Huanru et al. built a system named DeepJ which could compose music conditioned on a specific mixture of composer styles [15]. However, music generated through the above works still has the problem of lacking long-term structure and temporal correlation, which yet needs to be solved.

This work has improved the network structure on the basis of BALSTM, and introduced the thought of Lookback [6], strengthening the relation within bars and yielding better long-term structure. For emotional music generation tasks, we train the model with global condition of emotional vectors, and design tunable parameters for generating music of corresponding emotion. Subjective experiments show that people can distinguish different emotions expressed from most of the generated music samples.

II. Data Representation

A. Note Representation

The representation of a piece of music for T time steps and N pitches range is defined by matrix M as in (1). We take the time dimension as 128 and the pitch dimension as 48 (C3 to C6 for 4 octaves) in this work. A certain element at the position (t, n) of the matrix indicates the note data in n-th pitch at t-th timestep, and each note consists of three values: play, articulate, dynamic. Play and articulate are binary values (0 or 1) indicate if the note is being played and is being articulated or not. Dynamic is a continuous value which scaled between 0 and 1.

$$M_{T,N} = \begin{bmatrix} [1,0,0.5] & [0,0,0] & [0,0,0] & ... & [0,0,0] \\ [0,0,0] & [1,1,0.2] & [1,0,0.3] & ... & [0,0,0] \\ ... & ... & ... & ... & ... \\ [0,0,0] & [0,0,0] & [0,0,0] & ... & [0,0,0] \\ [0,0,0] & [0,0,0] & [1,0,0.3] & ... & [0,0,0] \end{bmatrix}_{T\times N} \quad (1)$$

B. Emotion Vector

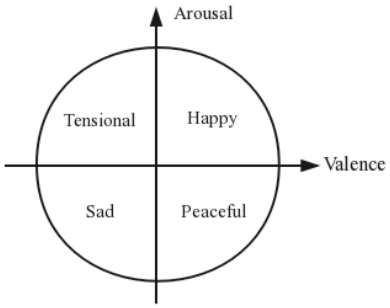

Fig. 1. Russell's 2-dimensional VA emotional space.

Russell expressed emotions as a 2-dimensional emotional space with valence (how active) and arousal (how positive) as two coordinate axes [16] as shown in Fig. 1. We regard the four quadrants as four basic emotions. We adopt one-hot coding to represent the four categories of emotions as [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]. In the machine learning algorithm, the distance is calculated in Euclidean space. This sparse representation method maps the discrete emotion classifications to Euclidean space, so that such representation is reasonable and effective.

III. Methodology

A. Neural Network Architecture

Extending beyond BALSTM [14], the model is introduced the thought of Lookback inspired by the project Magenta [6]. The overall architecture of the network is shown in Fig. 2 And we take advantage of parallelization in code shortening the time of model trained.

1) BALSTM

The BALSTM architecture generates polyphonic music by modeling each note within each time step as a probability conditioned on all previous time steps and all notes within the current time step that have already been generated, which is defined as:

$$P(X_{t,n} \mid X_{t,n-1}, X_{t,n-2}, ..., X_{t-1,N}, X_{t-1,N-1}, ..., X_{1,2}, X_{1,1}) \quad (2)$$

The network architecture is mainly composed of time-axis and note-axis which are shown in Fig. 2. After note data being processed and built as input layer, we connect it to time-axis layers. In time-axis (consists of two layers of stacked LSTMs), notes on all timesteps for each pitch are supplied to networks in recurrence. The output therefore is capable of remembering the time structure of music, it can be considered as high-level features that carry the temporal information feeding into note-axis. In note-axis (also consists of two layers of stacked LSTMs), it trains notes on all pitches for each time step recurrently, learning the vertical pitch relation among notes which is helpful for obtaining chord or counterpoint features of music. That's the reason why this architecture modeling polyphonic music along both two axes is so-called as Bi-axial LSTM (BALSTM). Both in time-axis and note-axis, the weights of each LSTM unit are shared across each note, forcing the networks to learn note invariant features.

2) Input Layer with LookBack

The input built for networks is a critical task for music generation. In the basic RNNs, it usually only takes note data as input. Moreover, some other note information is also constructed as input in BALSTM. On this basis, we further improve and enrich the input of network layers which is shown as Fig. 3.

- Note matrix: inspired by convolution in image recognition, the note data convolved through the a 2-octave window similar to a convolution kernel to attain transposition-invariance.
- Pitch class: it is a collection of notes that have a certain pitch on all octaves represented by the one-hot vector.
- Pitch position: We normalize the absolute pitch to the relative pitch position between 0 and 1.

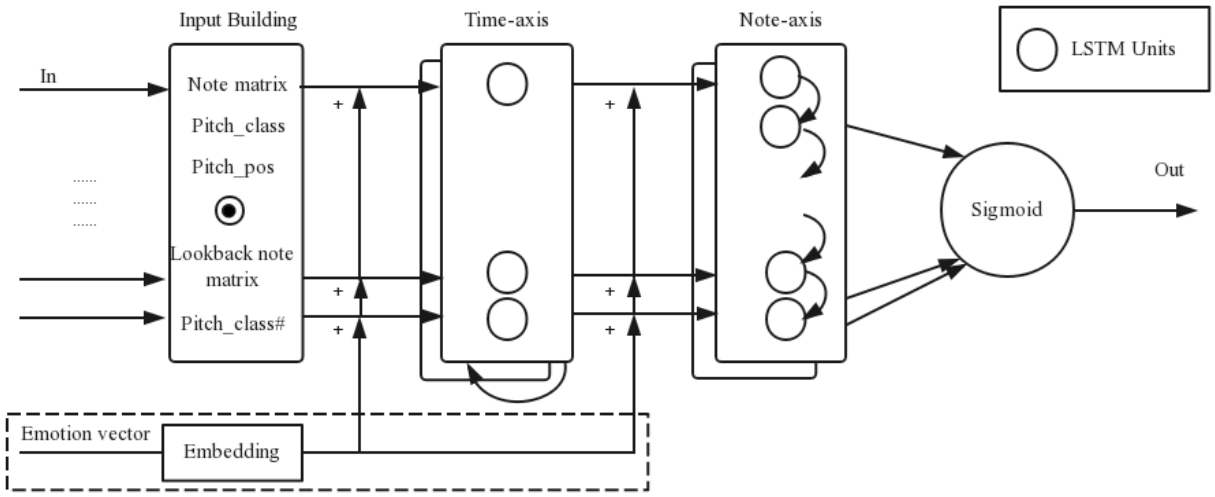

Fig. 2. The overall architecture in this work. (The dashed box section represents the additional portion in the task of emotional music generation; "◉" denotes the operation of concatenation , " + " denotes the operation of add.)

- Lookback note matrix: Lookback RNN was introduced to improve musical long-term structure. Compared with the basic RNNs, Lookback RNN not only input the previous note event, but also input the corresponding events one bar ago.
- Pitch class#: the number of played pitch class is also calculated as an input to complete the description of pitch class.

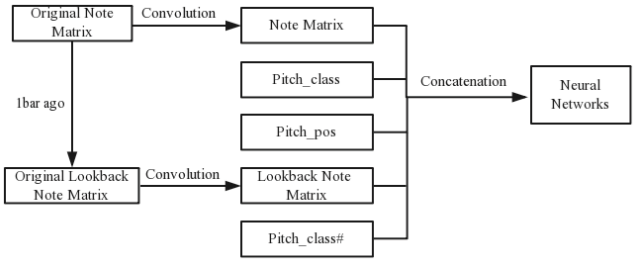

Fig. 3. Inputs of the networks.

All these parameters are built as an input through concatenation, and then feed in network layers for training.

B. Objective Function

The model produces three outputs for each time step at each pitch: the probabilities of play and articulate, and the values of dynamics. They are trained at the meantime. Play and articulate are trained as logic regression problems with log-likelihood functions, while dynamic is trained by mean squared error function. Formulas are shown below, where T is the number of timesteps and N is the number of possible notes. These three loss functions are combined as the final objective function:

$$L_{play} = \sum^{T,N} \ln[t_{play} y_{play} + (1-t_{play})(1-y_{play})] \quad (3)$$

$$L_{arti} = \sum^{T,N} \ln[t_{arti} y_{arti} + (1-t_{arti})(1-y_{arti})] \quad (4)$$

$$L_{dynamic} = \sum^{T,N} t_{play} (t_{dynamic} - y_{dynamic})^2 \quad (5)$$

C. Emotional Music Generation

1) Emotion Conditioning

In the task of generating emotional music by models, we need to learn the emotion information of music. Therefore, we adopt one-hot emotion vector just mentioned above as a conditioning information and add it as an additional input to the neural networks. As shown in the dashed box section in Fig. 2, we operate emotional vectors through embedding, and add them to time-axis layer and note-axis layer respectively in a manner of global conditioning, then all notes will share the same conditioning for each timestep.

2) Silence Correction

In the process of music generation, the generated samples may have longtime mute segments. To tackle this weakness, we set a temperature coefficient for generation. The output of network layers will be multiplied by the temperature, and the result will enter into sigmoid to calculate the final probability. When there is continuous mute in the sequence, the temperature will increase, which increases the probability to play the next note after sigmoid calculation. When non-mute note occurs, the temperature is reset to 1. So that it can avoid the generated samples having too much silence period.

3) Rhythm Control

We also give some rhythm control in the process of generating music corresponding to different emotions. While generating samples with high arousal emotion (happy and tensional), the resolution is slightly increased to speed up the rhythm; and generating music with low arousal emotion (sad and peaceful), it is adjusted to a relatively low rhythm resolution.

IV. EXPERIMENTS AND EVALUATIONS

A. Dataset

We trained our model on a widely used dataset: piano-midi. It includes piano pieces from 23 major classical composers which possesses higher complexity and richness compared to other MIDI datasets. We restricted the dataset to pieces with the 4/4 time signature and quantify a bar as 16 time steps (4 time steps per beat). For the emotion music generation task, we classified this dataset according to four basic emotions.

B. Training and Generation

At the training stage, dropout of 50% was applied to each LSTM layer as a regularizer to avoid overfitting, and Adam optimizer was used [18]. Two LSTM layers is used in the time-axis direction with 300 nodes each, and two LSTM layers in the note-axis direction with 150 nodes each.

$$Average\ NLL = -\frac{1}{S}\sum\frac{1}{TN}(L_{play} + L_{arti}) \tag{6}$$

We calculate the average negative log-likelihood (NLL) on the same test set of BALSTM and our model. The average NLL can evaluate the training quality of a model, which is defined as (6), where S is the number of samples. The NLL and training duration are shown in Table I. It can be seen that our model has better training effect and takes less time.

TABLE I. LOG-LIKELIHOOD PRFORMANCE AND TIME TRAINED

	NLL	Time trained
BALSTM (D.J.)	-4.90, -5.00	24-48
BALSTM +Lookback	-4.22, -4.43	12

At the generating stage, we set an optional parameter to decide how long the music to generate, which defaults to 32 bars. For the generation of emotional music, an emotional parameter is also set so that users can generate music in a specific emotion to meet their needs.

C. Evaluations

We conduct both subjective and objective experiment to evaluate the quality of music generated. What's more, in order to evaluate how well can our model create emotionally distinct music, we test if humans can identify the emotion of music generated.

1) Quality Evaluation

a) Musical Metrics

Quality evaluation of music was done using a number of musical measurements on generated output as in [12].

- **Polyphony (P)** measures how often (at least) two tones are played simultaneously (their start time is exactly the same)
- **Scale consistency (SC)** is calculated by counting the fraction of tones in all standard scales, and reporting the proportion for the best matching scale.
- **3-Tone Repetitions (3TR)** of music sequences are computed by counting how much 3-tone recurrence in a sample and reporting the proportion score of it in the whole piece. It takes only the order of tones into consider.
- **Tone span (TS)** is the number of half-tone steps between the lowest and the highest tone in a sample.

We compare the music samples generated from this work with those of the original model. And these metrics of the real music in the training set was also calculated as a reference.

As shown in Table II, the music generated in our work is closer to the real music sample in terms of the Polyphony and Scale consistency, and the 3-tone repetition is also better than the original BALSTM. To some extent, the 3-tone repetition and the scale consistency can reflect whether the music acquire a better long-term structure. Since the pitch range in this work is set to 48, the Tone span gets a lower point naturally.

TABLE II. MUSICAL METRICS FOR QUALITY EVALUATION

	P	SC	3TR	TS
BALSTM(D.J.)	0.5374	0.8140	0.1101	58.6
BALSTM +Lookback	0.3969	0.8265	0.2574	45.5
Real Music	0.3643	0.8268	0.6312	57.1

b) User Study

Music is an auditory art. In addition to the objective metrics, more importantly, we conduct a listening test to make the audience evaluate the quality according to their subjective feelings in the following. As in [17], we surveyed 30 subjects to grade the samples in terms of following aspects:

- **Harmonious(H):** if the samples have pleasant harmony,
- **Rhythmic(R):** if the samples have unified rhythm,
- **Musically structured (MS):** if the samples have clear musical structure,
- **Coherent(C):** if the samples are coherent.

We measure these parameters in a 5-point Likert scale. In the test, participants were given 10 music samples half generated by this work and half from original BALSTM. 30% of the subjects acquire professional music backgrounds, while the rest doesn't. The evaluation result is shown in Table III. The model in this work is preferred by professional group (pros) and non-professional group (non-pros) and performs better overall.

TABLE III. USER STUDY RESULTS FOR QUALITY EVALUATION

		H	R	MS	C	Overall
Non-pros	**BALSTM**	2.99	3.03	3.00	3.56	3.14
	BALSTM +Lookback	3.75	3.75	3.53	3.68	3.68
Pros	**BALSTM**	2.53	2.67	2.49	2.73	2.61
	BALSTM +Lookback	3.04	2.97	2.93	3.13	3.02

2) Emotion Classification

Similar to the human survey to analysis different styles of generated music in [15]. We invited 30 listeners for subjective evaluations of emotional classification.

Different music can arouse different emotions of listeners. However, it is not only determined by music itself, but also

related to the moods of listeners at that time. We generated 5 groups of music and 4 basic emotions for each group (a total of 20 pieces). In order to avoid subjective prejudice, the samples are messed up and then identified by the listeners. The accuracy of emotion classification is shown in Fig. 4.

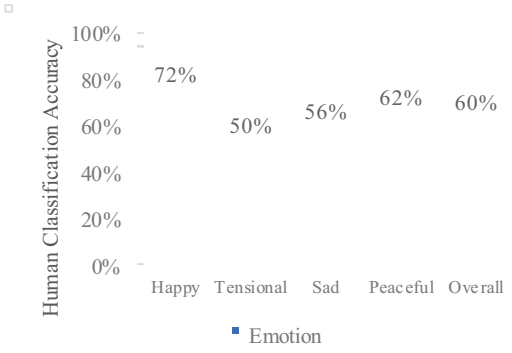

Fig. 4. Human accuracy on classifying samples into emotion categories

The music sample of Happy was correctly recognized with the highest accuracy of 72%, 62% for Peace secondly, while 56%, 50% for Sad and Tensional respectively. The overall accuracy is 60%. For this experiment, we also make statistical analysis counting all the option results as shown in Table IV.

TABLE IV. RESULTS OF HUMAN CLASSIFICATION

Generated samples \ Human classification	Happy	Tensional	Sad	Peace
Happy	72%	16%	2%	10%
Tensional	28%	50%	12%	10%
Sad	0	12%	56%	32%
Peace	2%	2%	34%	62%

It is found that active degree of emotion is easy to distinguish, whereas the positive degree relatively tends to be misjudged. Two emotions in similar level of arousal but opposite level of valence tend to be confused. The generated Happy music has 16% mistaken for Tensional emotion and 28% in the opposite case. The misjudgment between Sad and Peace are more obvious. The proportions of confusion between these two emotions even exceed 30%.

V. CONCLUSION AND FUTURE WORKS

In order to learn the relation within bars and obtain better long-term structure, we introduce the thought of Lookback into BALSTM structure. Through evaluation, it has been proved that the improved structure is superior to the original model. Furthermore, we applied this structure in the generation task of emotional music for the first time, and obtained preliminary achievements. However, there are still many problems to be solved for emotional music generation. For instance, the emotional color of music is closely related to chord progression, and the chord information can be extracted from music as a feature representing different emotions, which may be remained as one of the future works.

ACKNOWLEDGMENT

This research was supported by the NSFC grant (No.61501410 and No.61631016) and the Engineering Planning Project of Communication University of China grant (No. 2018XNG1809 and No. 3132017XNG1716).

* Jingling Wang is the corresponding author.

REFERENCES

[1] P.M. Todd, "A connectionist approach to algorithmic composition, computer," Computer Music Journal, vol. 13, issue 4, pp. 27-43, 1989.

[2] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, issue. 8, pp. 1735-1780, 1997.

[3] D. Eck, J. Schmidhuber, "Finding temporal structure in music: blues improvisation with lstm recurrent networks". Neural Networks for Signal Processing XII Proceedings of the IEEE Workshop, pp. 747-756, 2002.

[4] B.L. Sturm, J.F. Santos, O. Bental, and I. Korshunova, "Music transcription modelling and composition using deep learning," 1st Conference on Computer Simulation of Musical Creativity, 2016.

[5] J. Wu, C. Hu, Y. Wang, X. Hu, and J. Zhu, "A hierarchical recurrent neural network for symbolic melody generation," unpublished.

[6] W. Elliot, D. Eck, A. Roberts, and D. Abolafia, "Project Magenta: Generating longterm structure in songs and stories," 2016.

[7] H. Lim, S. Rhyu, K.Lee, "Chord Generation from Symbolic Melody Using BLSTM Networks," unpublished.

[8] H. Chu, R. Urtasun, and S. Fidler, "Song from pi: a musically plausible network for pop music generation," unpublished.

[9] N. Boulanger-Lewandowski, Y. Bengio, P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription," Proceedings of the 29th International Conference on Machine Learning, vol. 18, issue 13, pp. 3981-3991, 2012.

[10] S. Lattner, M. Grachten, and G. Widmer, "Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints, " unpublished.

[11] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach: a steerable model for bach chorales generation," Proceedings of the 34th International Conference on Machine Learning, pp. 1362-1371, 2017.

[12] O. Mogren, "C-rnn-gan: continuous recurrent neural networks with adversarial training," Constructive Machine Learning Workshop (CML) at NIPS 2016 in Barcelona, Spain, 2016.

[13] L.C. Yang, S.Y. Chou, and Y.H. Yang, "Midinet: a convolutional generative adversarial network for symbolic-domain music generation," the 18th International Society for Music Information Retrieval Conference), Suzhou, China, 2017.

[14] D.D. Johnson, "Generating polyphonic music using tied parallel networks," Computational Intelligence in Music, Sound, Art and Design, EvoMUSART 2017, 2017.

[15] H.H. Mao, T. Shin, and G. Cottrell, "Deepj: style-specific music generation," 2018 IEEE 12th International Conference on Semantic Computing, pp. 377-382, 2018.

[16] J.A. Russell, "A circumplex model of affect," J. Personality Soc. Psychology, vol. 39, issue 6, pp.1161–1178, 1980.

[17] H.W. Dong, W.Y. Hsiao, L.C. Yang, and Y.H. Yang, "Musegan: multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2018.

[18] D. Kingma, J. Ba, "Adam: a method for stochastic optimization," International Conference on Learning Representations, 2015.