International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-13

https://doi.org/10.70917/ijcisim-2026-1786
Article

The Impact and Integration Path of Artificial Intelligence Composition Technology on the Traditional Music Composition Mode

Zimu Mao[1] and Mengxin Mao[1,*]

[1] Department of Folk Music Performance (Yangqin), Tianjin Conservatory of Music, Tianjin, 300000, China
* Correspondence author: cynbelle530@163.com

Abstract: With the rapid development of deep learning technology, artificial intelligence composition technology has been able to generate musical works that are basically equal to traditional music composition in terms of sensory indicators, which has made a substantial impact on the main position of traditional music composition mode. In this paper, recurrent neural networks are used to learn the time-dependent features of musical sequences, and then generate adversarial networks to enhance the authenticity of the musical score, and express the music potential space in the form of probability distribution parameters. Rule-based algorithms are introduced to provide intervenable structural constraints, effectively generating scores with original musical style and rhythmic characteristics, and based on the above methodology, we explore the impact and integration possibilities of AI on the traditional mode of music creation. Experiments show that the PC and PI values of the AI composition method based on the RVAE-GAN model are improved by 8.57% and 5.89% compared with the Music Transformer model. The AI composition technique approaches the average level of traditional human work composition mode in terms of sensory acceptance and specific style maintenance. The above results confirm that the impact of AI on the traditional music composition mode is substantial, and the integration paths such as emotion-structure dual-drive and complementary style inheritance are feasible directions to cope with the impact.

Keywords: Recurrent Neural Network; Generative Adversarial Network; RVAE-GAN Model; AI Composition; Traditional Music Composition

1. Introduction

Music composition is a field that requires a high degree of specialized knowledge and skills, and only musicians with a solid foundation in music theory and composition experience can compose [1-2]. And with the rapid development of Artificial Intelligence (AI) technology, AI composition technology has become increasingly mature, which has had a significant impact on traditional music composition and triggered widespread concern in the music industry [3-4]. AI composition technology, refers to the process of automatically generating musical works by utilizing AI algorithms, in particular, machine learning and deep learning technologies [5-6]. Its core principle relies on training models through a large amount of data, enabling computers to simulate human compositions and generate works with musical structure, harmony, rhythm and other elements [7-9].

Traditional music creation has long been limited by human resources, material resources and other factors, it is difficult to realize large-scale music material processing and cross-genre integration, AI can make up for this shortcoming, its powerful computing power and data analysis capabilities, so that the breadth and depth of music creation can be comprehensively expanded [10-11]. With the help of AI technology, anyone who only needs to master simple operations can quickly generate a musical score with a complete structure by selecting appropriate parameters or inputting creative directions [12-14].

However, AI music composition is still in the exploratory development stage, and there are deficiencies in certain details, while facing a series of challenges such as ethical and emotional deficiencies [15-16]. In this context, mining the fusion path between AI composition and traditional music creation mode has become an important way to improve the current music creation, and only by realizing this fusion and deepening the collaboration between AI and human musicians can we explore the new paradigm of human-computer symbiosis and co-creativity at a broader level [17-20].

In order to deeply explore the impact of AI composition technology on the traditional music creation mode and the subsequent integration path, this paper designs and implements a set of combinatorial intelligent composition methods. A recurrent neural network is used as the core of music sequence feature extraction to capture the evolution pattern of music sequence features in the time dimension. On this basis, a generative adversarial network is introduced to discriminate the authenticity of the generated samples to enhance the naturalness of the musical score. A variational self-encoder is used to model the music potential space so that the model can maintain the features of the original training data. This paper also embeds a rule-based algorithm to impose constraints on the structural parameters of the score to realize the generation of canonical scores. A combined subjective and objective evaluation system is designed to verify the effectiveness of the model in generating musical scores. The differences between intelligent composition and traditional music creation are evaluated in five dimensions, such as pleasantness, innovation, and hierarchy, to provide data support for the design of the fusion path.

2. Intelligent Composition Based on Multimodal Neural Networks and Rule Algorithms

2.1. Recurrent Neural Networks

Recurrent Neural Network (RNN) is a model mainly used to process time-series data. It is mainly used in areas such as natural language processing, speech recognition, and music generation. Compared with other neural networks, RNN is unique in that it has a memory function, which can memorize previous input information and apply it to compute the current output, i.e., the current output is not only related to the current input, but also related to the previous input. The unfolding of an ordinary RNN in the time dimension is shown in Fig. 1.

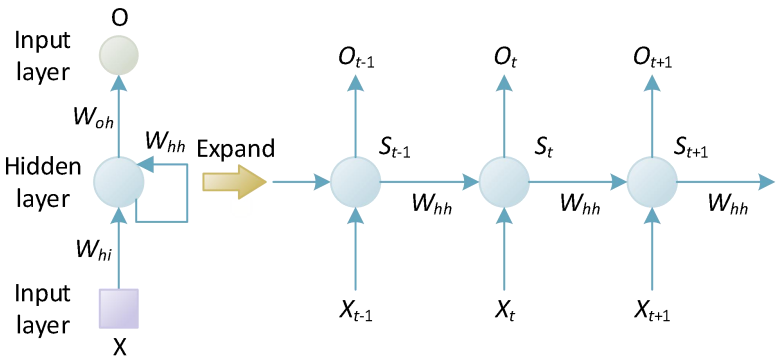

Figure 1. Ordinary RNN is unfolding in the time dimension

Figure 1 shows an RNN containing only one hidden layer, where X_t denotes the input at the t th moment, and similarly, X_{t-1} and X_{t+1} denote the inputs at the $t-1$ and $t+1$ moments, respectively, S_t denotes the state of the neurons in the hidden layer at the t th moment, W_{oh} is the weight between the hidden layer and the output layer, W_{hh} is the weight from the hidden layer to the hidden layer, and W_{hi} is the weight from the input layer to the hidden layer.

For any non-initial moment $t\left(t>1\right)$, the state of the hidden layer at that moment can be computed:

$$S_t=\sigma\left(W_{hi}X_{t-1}+W_{hh}+S_{t-1}+b\right) \tag{1}$$

Here σ is an activation function, commonly used activation functions are sigmoid, tanh and ReLU, and b denotes the bias. At the initial moment, the state of the hidden layer is generally set to zero.

The hidden layer to the output layer is calculated as:

$$o_t = \sigma\left(W_{oh}S_t + b\right) \tag{2}$$

2.2. GAN Adversarial Networks

Generative Adversarial Networks, or GANs for short, consist of two components: a generative model, and a discriminative model. The generative model, abbreviated as G, generates synthetic data and passes it to the discriminative model by modeling the joint probability and adding random noise. The discriminative model, abbreviated as D, is a binary classifier that aims to distinguish the data in the dataset from the data generated by the generator by modeling the conditional probability $P\left(Y\middle|X\right)$. Its entire framework is trained through neural network back propagation. Discriminative modeling also has a large application in the field of music recognition, where it serves to map high-dimensional feature vectors of music sequences into low-dimensional labels. A large number of music prior distributions and artificial music datasets are necessary elements for music generative models, and the prerequisite for modeling is to choose an appropriate prior distribution. In this paper, the generative model G aims to learn and approximate the distribution of real samples, with the goal of transforming the input random noise into music that is close to the real distribution, i.e., the more similar the generated music is to the music in the training set, the better. The discriminative model D is tasked with discriminating whether the input data is real or fake, and the training goal is to feed the generative model with information to help it generate more similar pieces of music.

GANs models are generally in the form of neural networks, and in this paper, we define a multilayer perceptron as the generator, which $G(z';\theta_g)$ is a parameterized differentiable function that learns the probability distribution X of the dataset through the defined noise variable $p_z(z)$ to form the probability distribution P_g and maps X is mapped into the data space. In this paper, we define a multilayer $LSTMD\left(x;\theta_d\right)$ whose output is a constant. $D(x)$ denotes the probability that x comes from the real data distribution instead of P_g. Training D maximizes the probability that the artificial and algorithmic scores correspond to the correct label. When training G, we need to minimize $\log\left(1-D\left(G\left(z\right)\right)\right)$, i.e., the more similar the data distribution of the result of the generative model G is to the distribution of the artificial music, the better. Based on this idea, we can use an optimization objective of the following form:

$$\begin{aligned}\min_G \max_D V(D,G) = {} & E_{x\sim p_{data}(x)}\left[\log D(x)\right] \\ & + E_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right]\end{aligned} \tag{3}$$

The above equation needs to be maximized when updating D and vice versa when updating G. When updating the discriminant model D, with respect to the sample x selected from the true distribution p_{data}, the closer the output of $D(X)$ is to 1, the better, and the larger $\log D(x)$ is, the better, i.e., the discriminator is able to recognize the difference between the true samples $\log D(x)$ and y=1 to form the loss function, which can then be backpropagate the gradient and thus update it. With the data $G(Z)$ generated by the noise Z, the closer $D(G(z))$ is to 0, the better, and the larger $\log\left(1-D(G(z))\right)$ is, the better, i.e., D can discriminate the false samples. And when updating the parameters of G again, $G(z)$ and the real data distribution should converge, when $D(G(z))$ is close to 1, then $\log\left(1-D(G(z))\right)$ is very small. When the global optimum is reached: $P_g = P_{data}$, at this point the loss and parameters of the generator G are retained and determined for the time being, and the discriminator D is considered. The optimal discriminator D is obtained as shown in equation (4):

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \tag{4}$$

Indeed, the optimization objective of the discriminator can be equated to the maximum likelihood estimation of the conditional probability $P\left(Y = y\middle|x\right)$, with x originating from P_{data} when $y = 1$,

and x originating from P_g when $y=0$. Thus the minimization-extremum problem can be deformed as:

$$\begin{aligned} Loss(G) &= \max_D V(G,D), = E_{x\sim p_{\text{data}}}\left[\log D_G^*(x)\right] \\ &+E_{z\sim p_z}\left[\log\left(1-D_G^*(G(z))\right)\right] \\ &= E_{x\sim p_{\text{data}}}\left[\log D_G^*(x)\right]+E_{x\sim p_g}\left[\log\left(1-D_G^*(x)\right)\right] \\ &= E_{x\sim p_{\text{data}}}\left[\log\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}\right]+E_{x\sim p_g}\left[\log\frac{p_g(x)}{p_{\text{data}}(x)+p_g(x)}\right] \end{aligned} \tag{5}$$

When and only when $p_g = p_{data}$, $Loss(G)$ reaches a global minimum, at which point $Loss(G)=\log\frac{1}{2}+\log\frac{1}{2}=-\log 4$. Therefore, when $p_g = p_{data}$, the model finally reaches convergence, at which time the data distribution of the generated samples' training samples' data distribution reaches basic convergence. Thus, the generated music and the real music to achieve the effect of fake to real.

2.3. VAE network

In this paper, we apply the Variational Autocoder (VAE), a generative network structure based on Variational Bayesian (VB) inference, as a method for studying music generation models. Unlike traditional self-encoders that describe the latent space by means of numerical values, it expresses the observation of the latent space in terms of probability distribution parameters, which has an irreplaceable contribution to the task of data generation.

VAE is a directed probabilistic model that consists of a network of encoders and a network of decoders.VAE and GAN serve the same purpose in music generation.

Implicit variables need to obey specific distributions to achieve better results. The KL scatter of the implied variables is framed to limit the value of the KL scatter so that the two distributions are as similar as possible.The KL scatter measures the size of the gap between the two distributions.Assuming that the probability density functions of the two musical distributions P and Q are p_x and q_x , the KL scatter between the two distributions is defined as follows:

$$D_{KL}\left(P\|Q\right)=\int_{-\infty}^{+\infty}p(x)\log\frac{p(x)}{q(x)}dx \tag{6}$$

In order to make the training results of the implied variables maximally conform to our pre-given distributions, we can make Q as the pre-given distribution such as the standard normal distribution, and P as the distribution of the implied variables to minimize the KL dispersion of both of them, at which point the P distribution will be asymptotically similar to the Q distribution during the training process.

In the VAE model, most of them will use the technique of reparameterization, in order to solve the problem of directly generating the distribution of implied variables, resulting in the computational map of the encoder and decoder can not be coherent to use the BPTT algorithm for the inverse computation. The so-called reparameterization means that in the actual computation, the encoder outputs parameters that the implied variables obey, and then these parameters further produce implied variables that obey a specific distribution. For example, if the normal distribution parameters are μ , σ , then the purpose of the encoder is to be used to generate the two parameters μ , σ . Knowing these two parameters, the VAE model needs to make the implied variables obey the $N(0,I)$ normal distribution. And based on these two parameters, the corresponding KL dispersion loss function KLLoss is calculated, by minimizing the KLLoss, the distribution of the implied variable and the target can be converged, that is, the standard normal distribution, this loss function is calculated by the probability density function of the normal distribution as follows:

$$KLLoss(\mu,\sigma)=-\frac{1}{2}\left(1+2\log\sigma-\mu^2-\sigma^2\right) \tag{7}$$

The autocoder has two parts of the loss function, one for the implied variables to approximate the normal distribution and one for the output data distribution to approximate the input data distribution,

which account for the weights of the hyperparameters. The cross entropy is calculated for each element then it is summed. The formula of cross-entropy and entropy is utilized to calculate the KLD, $q\left(z\middle|x^{i}\right)$ is the encoder and $f(z)$ is the decoder:

$$LOSS = KL(q,p) = H(q,p) - H(p) \tag{8}$$

$$\int p(x)\log\frac{1}{q(x)}dx = E_{P(X)}\left(\log\frac{1}{q(x)}\right) \tag{9}$$

$$H(p) = E_{P(X)}\left(\log\frac{1}{p(x)}\right) = -E_{p}(\log p) \tag{10}$$

$$H(q,p) = E_{P(X)}\left(\log\frac{1}{p(x)}\right) = -E_{p}(\log q(x)) \tag{11}$$

2.4. Multimodal neural network model

Combining the above two and the advantages and disadvantages of each network, we propose a new multimodal neural network model: rule-based neural network (RVAE-GAN). This network model includes an encoder E, a generator G (decoder) and a discriminator D, and the structure of the multimodal neural network model is shown in Figure 2.

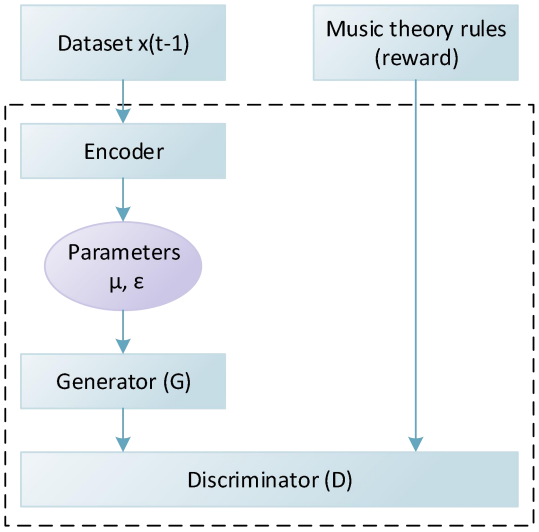

Figure 2. System diagram

In the VAE decoder and the GAN generator, they were merged into one by having them share parameters and trained together. The main architecture of all three networks used in this part incorporates a convolutional neural network. The input time-dependent data is treated as a series of individual frames with internal correlations. And these frames are generated using convolutional neural network extraction while maintaining their interdependencies. For each pair of consecutive frames, an encoder is used to encode the previous frames as their corresponding latent information. The generator then attempts to generate (predict) information for subsequent frames from the latent distribution of the previous frame. This combines the current information with the information from the previous frame to generate the next desired content. Each pair of current real training frames and synthesized frames is then forwarded to the discriminator as real and fake data, respectively.

2.5. Rule algorithm

The degree of fulfillment of a set of compositional rules or constraints formed by music theory to form the REWARD function. The melodic line is one of the most important means of expression in music, and in a general sense music cannot exist without it and rhythm. The expressive power and regularity of the melodic line is based, to a considerable extent, on the natural connection between the

upward progression of increasing tension and the downward progression of decreasing tension.

Here in this thesis, according to the constraints of the intervals in the melodic line, we try to avoid big jumps of more than five degrees, and the intervals in the melodic line avoid big jumps of more than an octave while avoiding consecutive big jumps in the same direction; in the ascending melodic line formed by the three notes in sequence, the big jumps are usually followed by a small interval in the descending melodic line formed by the three notes in sequence; and we try to avoid the use of triple whole tones or diminished fifths in the melodic line as well as repeating the same note in succession. If the interval difference between two neighboring notes is greater than an octave, it is noted as 0, or 1, and $g_1(x)$ is the average value for each piece of music. That is, if a piece of music has a groups of neighboring notes whose interval difference is greater than an octave, the remaining b groups are less than or equal to an octave. This can be expressed by the following formula:

$$g_1(x) = \frac{a*0 + b*1}{a+b} \tag{12}$$

For music, a pitch that goes up all the time or down all the time would not be good music. So, use $g_5(x)$ to evaluate the overall contour of a piece of music. The definition is as follows:

$$g_5(x) = \begin{cases} 1 & \frac{1}{n-1}\left|\sum_{i=1}^{n-1} interval_i\right| \le 1 \\ 0 & else \end{cases} \tag{13}$$

A pitch line is a line in which the pitch of a tone moves in time. The pitch line is dominated by cascades or small jumps, supplemented by large jumps. To guide the design of the algorithm, a model of the pitch rhythm rule is constructed (denoted as R_p):

$$\Delta p = \left| p(i) - p(i-1) \right| \in S_I \tag{14}$$

$$S_I = \left\{ I \mid I \ge 0, I \le I_{\max}, I \in Z \right\} \tag{15}$$

$$I_{\max} = \begin{cases} 5, & If\ p(i) \in S_p, p(i-1) \in S_p \\ 12, & If\ p(i) \in S_p, p(i-1) \notin S_p \end{cases} \tag{16}$$

$p(i)$ is the pitch of the ith note, S_I is the set of samples of the intervals of the neighboring notes, each sample value represents the number of semitones separating the two tones (intervals are all expressed in terms of semitones), and S_p is the set of samples of the pitches within a perceptual unity. Perceptual unity is the same structure within the human senses, if two tones are not in a perceptual unity, even if they are adjacent to each other, there will be no sense of unity in the judgment, or the sense of unity experience is weak. The musical message expressed by R_p is "use only cascades or small jumps within a perceptual unity, and large jumps between perceptual unities".

Consonant intervals (i.e., intervals with small common multiples of vibrational frequency ratios, e.g., pure 1st, 4th, 5th, and 8th octaves, and large and small 3rd and 6th octaves) are used as much as possible, and dissonant intervals (i.e., intervals with large common multiples of vibrational frequency ratios, e.g., large and small 2nd and 7th octaves) are used sparingly or not at all. However, the excessive use of harmonic intervals will produce a "single" and "un-rich" sensation in the hearing, and this applies to both harmonic and melodic intervals, because even after the physical stimulation of the sound has ceased, the auditory impression of the sound remains in the brain. This applies to both harmonic and melodic intervals. This consideration is also in contradiction with the pitch rule R_p. In order to take into account the beauty of the music and the object of expression as well as R_p, a "uniform rule" R_u is set here:

$$I_b \in \{0,2,3,4\} \tag{17}$$

$$I_{bb} \in \{0,2,3,4,5,7\} \tag{18}$$

$$I_{\to p} \in \{7,9,10,12\} \tag{19}$$

$$N_{i=1} \in \{1,2\} \tag{20}$$

where I_b, I_{bb}, and $I_{\to p}$ denote the values of the interval samples corresponding to the intervals within, between, and towards the climax of the measure, respectively, and $N_{i=1}$ is the number of minor second intervals (with a semitone number of 1) in a section, which is limited to be used no more than 2 times because of the strong tension of the interval.

The regular model R_d is constructed to represent the tempo and rhythmic changes of the piece:

$$DC(i) = D(i) / D(i-1) \in [2^{-a}, 2^{a}] \tag{21}$$

$$a = \begin{cases} 2, & D(i) \in S_p, D(i-1) \in S_p \\ 3, & D(i) \in S_p, D(i-1) \notin S_p \end{cases} \tag{22}$$

$$N_t \in \{1,2\} \tag{23}$$

$DC(i)$ is the contrast between the temporal values of neighboring notes, $D(i)$ is the temporal value of the i th note; N_t is the number of long-temporal notes within perceptual unity. A piece of music cannot vary too drastically, and $g_7(x)$ is used to evaluate this situation. The definition is as follows:

$$g_7(x) = \begin{cases} 1 & \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\left(\left|\text{int } erval_i\right| - \left|\overline{\text{int } erval}\right|\right)^2} \le 1 \\ 0 & else \end{cases} \tag{24}$$

For music, pitch going all the way up or all the way down is not going to be good music. So, use $g_5(x)$ to evaluate the overall contour of a piece of music. The definition is as follows:

$$g_5(x) = \begin{cases} 1 & \frac{1}{n-1}\left|\sum_{i=1}^{n-1} interval_i\right| \le 1 \\ 0 & else \end{cases} \tag{25}$$

If a piece of music continues on one pitch or time value, the listener will find it boring. Therefore, if more than four notes in a piece occur consecutively on the same pitch and time value, then this parameter will be notated as 0 and vice versa as 1.

The time value is the length of time or a specific length of interval. Since drastic changes in the time value of two consecutive notes can be irritating to the listener, this is noted as 0 if there is a change in the time value of more than 4, and vice versa.

Since music is generated randomly, there may be large intervals between consecutive notes. The maximum interval between notes can be specified, and when the interval exceeds the specified maximum value, lower REWARD.

Notes in strong beat position: In a measure, the first strong beat and the second strong beat are the two most significant beats in the measure. Positive REWARD is assigned when the first strong beat and the second strong beat are harmonics or rests, and no REWARD is assigned when they are not harmonics but have a scale note, or no scale note.

3. Experimental results and analysis

3.1. Preparation of the experiment

The experiments in this section are trained and tested using the Nottingham dataset, where the data is divided into three sets, the training, validation and test sets, containing 690, 177 and 170 tunes, respectively. In this paper, training experiments were first conducted for 24 hours using default hyperparameters with the number of iterations set to infinite.

During training, the model will adjust its parameters to make its output closer to the dataset after learning the data every time, so that the process of model self-adjustment is called a step. with the increase of the number of steps, the model will be gradually fitted to the dataset, and the generated music results will be very different. The results of the training experiments are shown in Figure 3. 8000steps of the model generated music is too fluttering, 34000steps of the model lacks of change and

hard articulation, and 15000steps of the model generates the most fluent and natural music. The music generated at 15000step is the most smooth and natural. In the early stage of migration training, the training loss fluctuates drastically, but after 15000 iterations, the loss is stabilized at 0.2 or less.

Therefore, the learning rate is changed from 1e-4 to 1e-5 to obtain a less fluctuating loss curve. The number of iterations was set to 15000 to avoid possible overfitting from subsequent iterations. The training loss versus step curve is shown in Figure 4, where a lower learning rate usually leads to slower convergence, but the fluctuations are still close to stabilizing within 0.2.

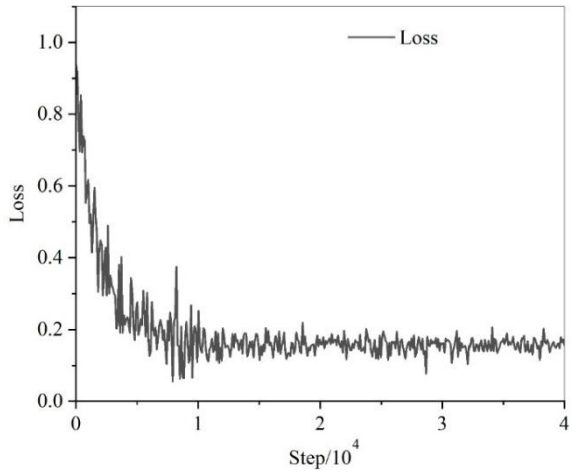

Figure 3. Training experiment

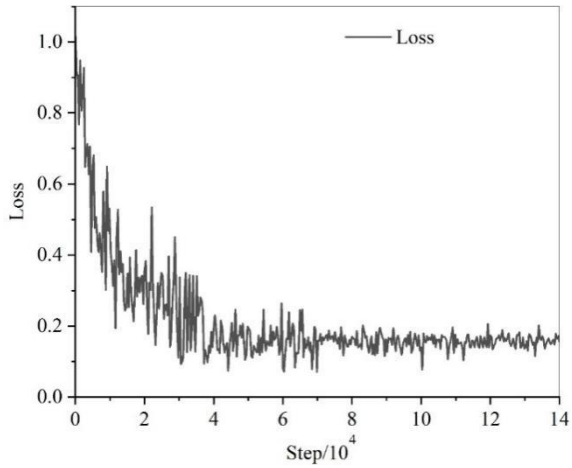

Figure 4. Training loss in migration learning

3.2. Objective evaluation

The Magenta project of Google Labs is an open-source tool organized by Google that specializes in the study of the artistic aspects of artificial intelligence. The objective evaluation of generated music in the Magenta project is mainly conducted from the aspects of tempo, melody and completeness, and evaluated by questionnaires and scores. The evaluation of the generated music also mostly uses manual evaluation to test the listening sensation of the music from more subjective perspectives such as rhythm, melody and integrity, and is evaluated by questionnaire and scoring.The team of MuseGAN proposes a numerical evaluation criterion that can evaluate the generation effect of multi-track music, but only applies to the music in midi format. This section draws on MuseGAN's criteria and designs 2 objective evaluation metrics.

The correctness (PC) metric is designed to detect whether the model can “understand” the grammar well, which is calculated as shown in Equation (26):

$$PC = \frac{\text{Number of tracks with more than 4 bars}}{\text{Total number of tracks generated}} \times 100\% \tag{26}$$

Drawing on the QN metrics, a completeness indicator (PI) is set to detect whether the model's

overall music generation is too fragmented.

$$PI = \frac{\text{Number of tracks with more than 4 bars}}{\text{Total number of tracks generated}} \times 100\% \tag{27}$$

The Folk-RNN, Char-RNN, Music Transformer and RVAE-GAN models are compared in the experiments, and all three models are trained with the same training set, the generation length is set to 512 bytes, and 150 segment samples are generated for comparison, and each group of experiments is repeated 10 times, and the average value is taken.

The comparison results are shown in Table 1. In terms of PC and PI, RVAE-GAN and Music Transformer are significantly stronger than Folk-RNN and Char-RNN models, and RVAE-GAN is stronger than Music Transformer, with PC and PI values of 89.13%, 98.15%, and 98.15%, respectively. The PC and PI values of RVAE-GAN are 89.13% and 98.15%, which are 8.57% and 5.89% higher than those of Music Transformer.

Table 1. The effects of the music are compared

Model	PC/%	PI/%
Folk-RNN	53.51	63.20
Char-RNN	74.42	82.45
RVAE-GAN	89.13	98.15
Music Transformer	80.56	92.26

3.3. Subjective evaluation

In this paper, we designed a human evaluation experiment to evaluate its quality, using the same metrics as a band, and demonstrated its polyphony generation ability together for easy comparison. C-RNN-GAN, MusicVAE, and LSTM-VAE-GAN models are added for comparison, and the same training, validation, and test sets are used as those used to train the RVAE-GAN model. The metrics are expressed as:

Rhythm: does the music sound smooth and are the pauses appropriate?

Melody: are the note relationships natural and harmonious?

Completeness: is the music structurally complete without sudden interruptions?

Polyphony: is it possible to generate polyphonic music.

Fifteen pieces of music were randomly selected from each of the model's generation results and randomly numbered, and 15 pieces of music were also randomly selected from the test set to serve as a control. Forty testers were invited to score each piece of music on the above three items of rhythm, melody, and completeness, with scores ranging from 1 to 5, and the average of the scores of each modeled piece of music was counted, and the results of the test are shown in Table 2.

According to the test results, the score of the RVAE-GAN model has been greatly improved compared with several other types of models, and has even been closer to the real music. The average score of the subjective generation quality of the RVAE-GAN model is 3.72, and the difference is only 0.03 points compared with the average score of the subjective generation quality of the real music, which proves the effectiveness of the RVAE-GAN model in music generation. A better listening experience can also be obtained using the Char-RNN model, which consists of only RNNs. Among the monophonic music generation models, the LSTM-VAE-GAN model scores slightly lower than Music VAE, but its auditory integrity is greatly improved, and it is hypothesized that the LSTM-VAE structural model can improve the integrity of the generated sequences to a certain extent when nested with the GAN model.

Table 2. Subjective generated quality test results

Test sample source	Compound tone	Rhythm	Melody	Integrity	Mean
Nottingham	YES	3.77	3.72	3.77	3.75
RVAE-GAN	YES	3.74	3.63	3.79	3.72
Char-RNN	YES	3.47	3.37	3.4	3.41
LSTM-VAE-GAN	NO	3.02	2.89	3.41	3.11
C-RNN-GAN	NO	2.97	2.89	2.57	2.81
Music VAE	NO	3.08	3.19	3.26	3.18

During the testing process, five pieces of music were taken from each model generation result and each test set, and the testers were informed that the music they were about to hear was partly composed

by humans and partly model generated, and they were asked to judge whether the music they heard was model generated, and the number of people who judged the music to be model generated was counted according to each piece of music, and the results of the test are shown in Table 3.

From the test results, it can be seen that even if the music is composed by human beings, it is still considered to be model-generated, and the RVAE-GAN model can be closer to the test set in the results, and the difference between its Turing test score and the test set score is only 0.2, and the volunteers can't distinguish the model-generated music from the human-composed music very well.

Table 3. Music test results

Test sample source	Song 1	Song 2	Song 3	Song 4	Song 5	Mean
Nottingham	16	17	9	13	6	12.2
RVAE-GAN	10	15	8	13	16	12.4
Char-RNN	13	16	22	17	22	18.0
LSTM-VAE-GAN	18	26	18	19	22	20.6
C-RNN-GAN	25	22	27	23	21	23.6
Music VAE	20	23	23	16	24	21.2

3.4. Comparison of Intelligent Composition and Traditional Music Composition

The 36 pieces of music randomly selected for this experiment were 12 folk songs composed by artificial composers, 12 composed by the design model in this paper, and 12 composed by the basic_rnn model in the Magenta project of Google Lab. The five parameter indexes of pleasantness, innovation, hierarchy, concordance and style clarity were selected to compare the intelligent composition and traditional music creation model. because the five parameter indexes are equally important, the scoring results of all the people on these music indexes were averaged first to get the average score of the five parameter indexes, and then compared separately. basic_rnn algorithm composition and this paper algorithm composition comparison results As shown in Figure 5, the comparison results of the 5 indicators scores of this paper's model compositions and people's compositions are shown in Figure 6.

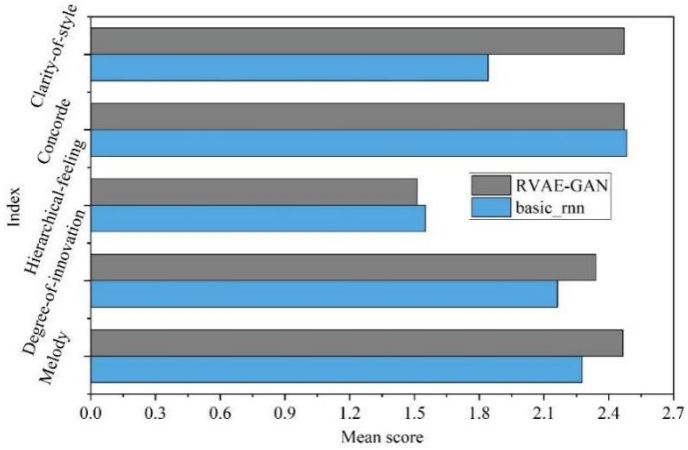

Figure 5. The basic_rnn algorithm is compared with the algorithm in this article

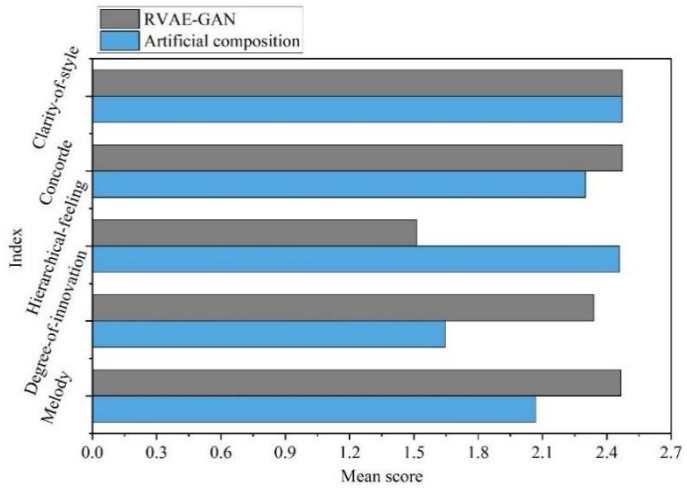

Figure 6. This paper compares the five indicators of artificial composition

The pleasantness and concordance of the music generated by the model of this paper and the music generated by the model of the basic_rnn algorithm are basically the same as those of the human work song, which indicates that the music can basically satisfy people's needs from the sensory point of view. From the point of view of style clarity, the model of this paper generates higher music than the model of basic_rnn algorithm, because the model of basic_rnn algorithm adopts basic one-hot coding to extract melodic features as the neural network input, while the variational self-encoder adopted in this paper expresses the observation of potential space in the form of probability distribution parameter, which makes the generating music retain the rhythmic style characteristics of folk songs better. the rhythmic style characteristics of folk songs. The structure of the basic_rnn algorithm model is more suitable for generating Western music. Generated music in terms of innovation, the network designed in this paper consists of a combination of algorithms, recurrent neural network learns the characteristics of the music from the training samples, and the generated motivic melody introduces a certain degree of randomness and flexibility. This makes the output music neither too random nor dull. However, from the point of view of hierarchy, the music generated by the model in this paper has a large gap with the human work piece, and the hierarchy of the music also affects the quality of the music, and there is still room for further improvement in the model designed in this paper.

From the above results, it can be seen that the AI composition technology has approached or even exceeded the benchmark level of some traditional human composition models in terms of sensory acceptance and the ability to maintain specific styles. This progress has caused a multifaceted impact on the traditional music composition mode.

In terms of the subjectivity of creation, the central position of the composer in the traditional mode of composition has been shaken, and music creation has been transformed from a spiritual output from the inside out to a calculable technical process.

In terms of creation techniques, traditional composition relies on the creator's in-depth understanding of a certain musical style, while models based on AI composition technology do not need to learn these rules, and can better retain the rhythmic style of the song through probability distribution modeling alone. At the level of creative efficiency, AI-based music creation models can batch generate music fragments that are not significantly different from human works, which puts traditional creation models at risk of being replaced. How to redefine the identity of composers and the boundaries of human-computer collaboration under the impact of the above is precisely the question that needs to be answered nowadays.

4. The integration path of AI composition and traditional music creation

Most of the existing AI composition systems have relatively low overall intelligence, mostly based on built-in MIDI music signals for machine learning and creation, lacking the human recognition system of musical emotions and without anthropomorphic music composition thinking. The human-computer interaction system is also limited to surface information exchange, and the machine performs the corresponding tasks in a passive form according to the user instructions obtained from the surface information exchange. The core direction of the future fusion path between AI composition and traditional music creation lies in the following: take machine vision, machine hearing and other multi-channel intelligent information fusion as the technical basis, and build a human-computer

interaction intelligent composition system that can understand human music emotion expression. One of the integration paths is to establish a dual-driven model of “human beings set emotional goals and structural frameworks, and AI generates suitable materials”, so as to make AI become a conversational creative partner. The second integration path is to use AI as a tool to assist in recording and style modeling, such as using variable self-coding to extract style parameters such as rhythm and accent, and then in the creation stage, AI generates new material that matches the style for the inheritor to use.

In this social group, the existence of anything must comply with social behavioral norms and ethical principles, AI brings us a turnaround in the industry but also hides some ethical issues: AI replaces a large number of laborers, but what should happen to the replaced staff? Since AI can compose music in bulk, do we still need to train composers? When we want to listen to and study traditional music, AI can provide accurate services as fast as possible, so do we still need to travel to seek help from traditional artists? These questions are worth pondering. In the face of the complicated issues behind the use of AI technology, we should first improve the social responsibility and legal awareness of scientists from the source, and avoid developing technology products that threaten the security of human society.

In addition, the government should work together with the technology sector to establish a relevant management organization, set up a reasonable attribution mechanism, clarify the attribution of responsibility, as well as predict and curb the problems that AI will face in the future development.

5. Conclusion

In this paper, we introduce RVAE-GAN, a model for generating music sequence data based on artificial intelligence composition techniques.The model consists of an encoder, a generator, and a discriminator, which utilizes a convolutional neural network and constraints to learn the spatially localized correlations of the data in each frame. Frame-by-frame sampling is achieved by using the encoder to obtain the latent distribution of the following frame based on the previous frame. The final music sequence generated has a more consistent frame-to-frame consistency. Compared with existing models such as Folk-RNN and Char-RNN, the AI composition experiments based on the RVAE-GAN model in this thesis have better results, and RVAE-GAN outperforms the Music Transformer model, which is the next best model, and the PC and PI values of the RVAE-GAN model are higher than those of the Music The PC and PI values of RVAE-GAN model are improved by 8.57% and 5.89% respectively compared to the Music Transformer model, which makes the automatic music generation better.

Five parameter indicators, namely, pleasantness, innovation, hierarchy, concordance and stylistic clarity, were selected to compare the multidimensional differences between the intelligently generated music and the traditional music creation model through the experimental method of manual scoring. On this basis, the integration path of the two from impact to synergy is further explored. The fusion of AI composition and traditional music composition includes: a collaborative composition model with dual-driven emotion-structure, the design of AI recording and style modeling tools, and an interactive iterative composition model.

The experimental samples in this paper are more limited, and the interpretability issues of emotion modeling and aesthetics of AI-generated music can be explored in the future. The integration of AI compositional techniques with traditional music creation models is still at an early stage, but the basic direction is clear, i.e., instead of replacing people with algorithms, algorithms are used to expand the boundaries of human creative possibilities.

About the Author

Zimu Mao was born in Tongling, Anhui, P.R. China, in 2004. I am currently a senior undergraduate student majoring in Music Performance (Yangqin) at Tianjin Conservatory of Music in China, and have been admitted to pursue a Master's degree at Hong Kong Baptist University upon graduation. My main research direction is Yangqin performance and music education.

Mengxin Mao was born in Tongling, Anhui, P.R. China, in 1994. I earned my Ph.D. from St. Paul University. I am currently a lecturer at Tongling University. My primary research interests are yangqin performance and music education.

References

1. Jarrett, S., & Day, H. (2024). Music composition for dummies. John Wiley & Sons.
2. Hogenes, M., Oers, B. V., & Diekstra, R. F. (2014). Music composition in the music curriculum. US-China Education Review A, 4(3), 149-162.

3. Hart, B., & Andrews, L. (2024). AI-Powered Neural Networks for Music Composition and Generation. American Journal of Artificial Intelligence and Neural Networks, 5(6), 12-16.
4. Liu, C. H., & Ting, C. K. (2016). Computational intelligence in music composition: A survey. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1), 2-15.
5. Hernandez-Olivan, C., & Beltran, J. R. (2022). Music composition with deep learning: A review. Advances in speech and music technology: computational aspects and applications, 25-50.
6. Tan, X., & Li, X. (2021, October). A tutorial on AI music composition. In Proceedings of the 29th ACM international conference on multimedia (pp. 5678-5680).
7. Yang, W., Shen, L., Huang, C. F., Lee, J., & Zhao, X. (2024). Development status, frontier hotspots, and technical evaluations in the field of ai music composition since the 21st century: A systematic review. IEEE Access, 12, 89452-89466.
8. Deruty, E., Grachten, M., Lattner, S., Nistal, J., & Aouameur, C. (2022). On the development and practice of AI technology for contemporary popular music production. Transactions of the International Society for Music Information Retrieval, 5(1).
9. Pereverzeva, M. V. (2021). The prospects of applying artificial intelligence in musical composition. Russian Musicology, (1), 8-16.
10. Tigre Moura, F., & Maw, C. (2021). Artificial intelligence became Beethoven: how do listeners and music professionals perceive artificially composed music?. Journal of Consumer Marketing, 38(2), 137-146.
11. Zhang, Y. (2024). Utilizing computational music analysis and AI for enhanced music composition: exploring pre-and post-analysis. Educational Administration: Theory and Practice, 30(5), 269-282.
12. Zhao, H., Min, S., Fang, J., & Bian, S. (2025). AI-driven music composition: Melody generation using Recurrent Neural Networks and Variational Autoencoders. Alexandria Engineering Journal, 120, 258-270.
13. Gioti, A. M. (2020). From artificial to extended intelligence in music composition. Organised Sound, 25(1), 25-32.
14. Oh, H. S. (2024). Is ai music beautiful? a study of the ai composition model evom. International Review of the Aesthetics and Sociology of Music, 55(1), 139-158.
15. RaNa, M. R. H. (2025). The Impact of Artificial Intelligence on Music Composition and Performance. International Journal of Technology, Management and Humanities, 11(01), 17-34.
16. Rohrmeier, M. (2022). On creativity, music's AI completeness, and four challenges for artificial musical creativity. Transactions of the International Society for Music Information Retrieval, 5(1).
17. Nikrang, A., & Kiesenhofer, S. (2025, September). Human–AI Co-Creation in Contemporary Composition: Interaction and Artistic Strategies with Ricercar. In Proceedings of the Conference on Animation and Interactive Art (pp. 65-73).
18. Ke, X., Xu, B., & Xie, Z. (2026). Beyond Composer Labels: Aesthetic Awe in Music Composed by Humans and AI. Computers in Human Behavior, 109020.
19. Zulić, H. (2019). How AI can change/improve/influence music composition, performance and education: three case studies. INSAM Journal of Contemporary Music, Art and Technology, (2), 100-114.
20. Raza, A. (2025). The evolution of human-AI collaboration in creative industries: case studies in music and design. Multidisciplinary Research in Computing Information Systems, 5(2), 105-123.