applied sciences

MDPI

Article

BoYaTCN: Research on Music Generation of Traditional Chinese Pentatonic Scale Based on Bidirectional Octave Your Attention Temporal Convolutional Network

Fanzhi Jiang [1], Liumei Zhang [1,*], Kexin Wang [2], Xi Deng [1] and Wanyan Yang [1]

[1] School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
[2] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: zhangliumei@xsyu.edu.cn; Tel.: +86-180-9233-0186

Abstract: Recent studies demonstrate that algorithmic music attracted global attention not only because of its amusement but also its considerable potential in the industry. Thus, the yield increased academic numbers spinning around on topics of algorithm music generation. The balance between mathematical logic and aesthetic value is important in music generation. To maintain this balance, we propose a research method based on a three-dimensional temporal convolutional attention neural network. This method uses a self-collected traditional Chinese pentatonic symbolic music dataset. It combines clustering algorithms and deep learning-related algorithms to construct a three-dimensional sequential convolutional generation model 3D-SCN, a three-dimensional temporal convolutional attention model BoYaTCN. We trained both of them to generate traditional Chinese pentatonic scale music that considers both overall temporal creativity and local musical semantics. Then, we conducted quantitative and qualitative evaluations of the generated music. The experiment demonstrates that BoYaTCN achieves the best results, with a prediction accuracy of 99.12%, followed by 3D-SCN with a prediction accuracy of 99.04%. We have proven that the proposed model can generate folk music with a beautiful melody, harmonious coherence, and distinctive traditional Chinese pentatonic features, and it also conforms to certain musical grammatical characteristics.

Keywords: music generation; pentatonic scale; clustering; 3D-SCN; BoYaTCN

Citation: Jiang, F.; Zhang, L.; Wang, K.; Deng, X.; Yang, W. BoYaTCN: Research on Music Generation of Traditional Chinese Pentatonic Scale Based on Bidirectional Octave Your Attention Temporal Convolutional Network. *Appl. Sci.* **2022**, *12*, 9309. https://doi.org/10.3390/app12189309

Academic Editor: Mauro Castelli

Received: 11 August 2022
Accepted: 13 September 2022
Published: 16 September 2022

1. Introduction

During the past few decades, the field of computer music has precisely addressed challenges surrounding the analysis of musical concepts [1,2]. Indeed, only by first understanding this type of information can we provide more advanced analytical and compositional tools, as well as methods to advance music theory [2]. Currently, literature on music computing and intelligent creativity [1,3,4] focuses specifically on algorithmic music. We have observed a notable rise in literature inspired by the field of machine learning because of its attempt to explain the compositional textures and formation methods within music on a mathematical level [5,6]. Machine learning methods are well accepted as an additional motivation for generating music content. Instead of the previous methods, such as grammar-based [1], rule-based [7], and metaheuristic strategy-based [8] music generation systems, machine learning-based generation methods can learn musical paradigms from an arbitrary corpus. Thus, the same system can be used for various musical genres.

Driven by the requirement for widespread music content, more massive music datasets have emerged in the genres of classical [9], rock [10], and pop music [11], for instance. However, a publicly available corpus of traditional folk music seems to pay little attention to the niche corner. Historically, research investigating factors associated with music composition from large-scale music datasets has focused on deep learning architectures, stemming from its ability to automatically learn musical styles from a corpus and generate new content [5].

Although music possesses its special characteristics that distinguish it from text, it is still classified as sequential data because of its temporal sequential relationship. Hence, recurrent neural networks (RNN) and its variants are adopted by most music-generating neural network models that are currently available [12–17]. Music generation sequence models were often characterized by the representation and prediction of a number of events. Then, those models can use the conditions formed by previous events to generate the current event. MelodyRNN [18] and SampleRNN [19] are representatives of this approach, with the shortcoming that the generated music lacks segmental integrity and a musical recurrent structure. Neural networks have studied this musical repetitive structure, called translation invariance [20]. Convolutional neural network (CNN) has been influential in the music domain, stemming from its excellence in the image domain. This regional learning capability is sought to migrate to the translational invariance of the musical context. Some representative work has emerged [21–23] to use deep CNN for music generation, although there have been few attempts. However, it seems to be more imitative than creative in music, stemming from its over-learning of the local structure of music. Therefore, inspired by whether it is possible to combine the advantages of both structures, they used compound architectures in music generation research [12,24–26].

Compound architecture combines at least two architectures of the same type or of different types [5] and can be divided into two main categories. Some cases are homogeneous composite architectures that combine various instances of the same architecture, such as the stacked autoencoder. Most cases are heterogeneous compound architectures that combine various types of architectures, such as a RNN Encoder-Decoder that combines the RNN and autoencoder. From an architectural point of view, we can conduct compositing using different methodologies.

- *Composition*—Combination of two architectures of the same type or of different types. For instance, the bidirectional LSTM [15] combines two RNNs to analyze music semantic contexts from temporal forward and inverse; and RNN-RBM architectures combine RNN architectures and RBM architectures [14].
- *Refinement*—Refinement and specialization of a model by additional constraints. The sparse autoencoder architecture is an example of a specialized solution to the note sparse coding problem on top of the autoencoder architecture [27] and the variational autoencoder (VAE) [28].
- *Nesting*—Nesting one model into another structure to form a new model. Examples include stacked autoencoder architectures [29] and RNN encoder-decoder architectures, where two RNN models are nested in the encoder and decoder parts of an autoencoder, so we can also call them autoencoders (RNN, RNN) [16].
- *Instantiation*—The architectural pattern is instantiated into a given architecture. For a case in point, the Anticipation-RNN architecture instantiates a conditional reflection architectural pattern onto an RNN and the output of another RNN as a conditional reflection input, which we can call conditional reflection (RNN, RNN) [17]. The C-RBM architecture is a convolutional architectural pattern instantiated onto an RBM architecture, which we can note as convolutional (RBM) [30].

Despite the success of each of these compound architectures in various periods, these works lack a wholeness in terms of generated musical segments (or musical translation invariance) and a certain creativity (or more prominent aesthetic value) in terms of musical semantics. In addition, the limitation of the dataset results in a rare research work in traditional Chinese folk music generation. For the above reasons, we investigate in this paper a symbolic domain generation architecture based on CNN and BiLSTM, focusing on the generation of traditional Chinese pentatonic folk music.

This research aims to develop a music generation model appropriate for traditional Chinese pentatonic folk music with near complete spatial structure (with musical transposition invariance), as well as reasonable temporal semantics (with musical genre identification). For this purpose, we first collected a folk music corpus on which to perform data analysis. Then, we performed a clustering-based genre analysis, resulting in a tra-

ditional Chinese pentatonic scale music dataset (TCPS). Afterward, we propose a neural network architecture based on CNN and BiLSTM, called BoYaTCN, which is a framework for symbolic music generation. The features of the musical data are first extracted by a convolutional neural network with a multilayer attention mechanism and a residual mechanism. Moreover, three-dimensional training is implemented by our described three-dimensional BiLSTM-based 3D-SCN architecture to learn music structure features from space and time. Furthermore, we conducted ablation experiments to assess the effects of the attention mechanism and the residual mechanism on the training process of music data, as well as to efficiently understand the interactions of music corpus attributes. This has potential implications for improving the accuracy of the melodic prediction and musical structure control. Subsequently, we conducted quantitative and qualitative analyses, and the experimental results indicated the model has good generalization performance. It can create traditional Chinese pentatonic scale music with specific features in a safe manner, which aids in understanding the intricate mechanism linking algorithmic music and genre culture, and positively encourages the preservation of traditional Chinese folk music.

The main contributions of this paper are summarized as follows:

1. A dataset of traditional Chinese folk music (TCPS) was created. To the best of our knowledge, this is the first publicly accessible dataset of traditional Chinese pentatonic scale folk music in MIDI format.
2. Driven by the motivation of genre-specific music generation, we conducted data mining on music data. Text clustering is applied into the field of symbolic music, which successfully identifies the stylistic features of music as a support for our genre-specific music generation. This is a novel proposal.
3. We proposed an architecture called 3D-SCN, followed by an architecture BoYaTCN based on compound neural network architecture, residual mechanism, and music attention mechanism for the generation of Chinese pentatonic scale music. The difficulty of balancing mathematical accuracy and aesthetic metrics of music creation is addressed, which is difficult to do with the present approaches.
4. We conducted experiments on our own established Chinese traditional folk music dataset and achieved encouraging results. Quantitative and qualitative evaluations demonstrate that our presented 3D-SCN and BoYaTCN have superiority in generating traditional Chinese pentatonic scale music.

The remainder of this paper is organized as follows. Section 2 presents the related work of this paper. Section 3 introduces some preliminary model details. Section 4 describes the establishment details of the dataset and conducts data-driven music genre clustering analysis. Section 5 states the details of the proposed model, 3D-SCN and BoYaTCN. Section 6 provides the data representation and detailed model setting during model implementation. Section 7 introduces objective metrics for model evaluation. Experiments and results are discussed in Section 8. Finally, Section 9 concludes the paper.

2. Related Work

In this section, we will describe the related work from two perspectives, based on the specificity of our work. First, we will investigate deep learning-based music generation architectures from three general branches, including RNN, CNN, and compound architectures. Moreover, research on Chinese folk music genre-driven music computing will also be surveyed.

2.1. Deep Learning-Based Music Generation

1. *RNN-based music generation*. This work [12] is an RNN architecture with a hierarchy of cyclic layers that generates not only melodies but also drums and chords. The model [13] well demonstrates the ability of RNNs to generate multiple sequences at the same time. However, it requires prior knowledge of the scale and some contours of the melody to be generated. The results demonstrate that the text-based Long short-term Memory (LSTM) performs better in generating chords and drums.

MelodyRNN [18] is probably one of the best-known examples of neural networks generating music in the symbolic domain. It includes three RNN-based variants of the model, two variants aimed at musical structure learning, lookback RNN and Attention RNN. Sony CSL [31] proposed DeepBach, which could specifically compose a polyphonic four-part choral repertoire in the style of J. S. Bach. It is also an RNN-based model that allows the execution of user-defined constraints, such as rhythms, notes, parts, chords, and allegro. However, this direction remains challenging for the following reasons. From the outside, the overall musical structure seems to have no hierarchical features, and parts do not have a unified rhythmic pattern. Musical features are considered extremely simplified in terms of musical grammar, ignoring key musical features such as note timing, tempo, scale, and interval. Regarding the connotation of music, the music style is uncontrollable, the aesthetic measurement is invalid, and there is a significant gap between the sense of hearing and the music created by the musicians.

2. *CNN-based music generation*. Some CNN architectures have been identified as an alternative to RNN architectures [21,22]. The paper [21] is proposed as a representative work for CNN-based generative model building, which enables speech recognition, speech synthesis, and music generation tasks. WaveNet architecture presents a number of causal convolutional layers, somewhat similar to recurrent layers. However, it has two limitations: its inefficient computation reduces the use of real time, and it was created to be mainly oriented to acoustic data. MidiNet [22] architecture for symbolic data is inspired by WaveNet. It includes an adjustment mechanism that incorporates historical information (melody and chords) from previous measurements. The authors discuss two ways to control creativity and constrain the condition. One method is to insert adjustment data only in the middle convolutional layers of the generator architecture. The other method is to reduce the value of two control parameters of feature-matching regularization, thus reducing the distribution of actual and generated data.
3. *Compound architecture-based music generation*. Bretan et al. [32] implemented the encoding of musical input by developing a deep autoencoder and reconstructed the input by selecting from a library. Subsequently, they established a deep-structured semantic model DSSM combined with LSTM to perform unitary prediction of monophonic melody. However, because of the limitations of unitary prediction, the quality of the generated content is sometimes poor. Bickerman et al. [24] proposed a music-coding scheme for learning jazz using deep belief networks. The model can generate flexible chords of different tones. It demonstrates that if the jazz corpus is large enough to generate chords, there is reason to believe that more complex jazz corpora can be performed. While some interesting pieces of jazz melodies have been created, the phrases generated by the model are not sufficient to represent all the features of the jazz corpus. Lyu et al [11] combined the capability of LSTM in long-term data training and the advantages of the Restricted Boltzmann Machine (RBM) in high-dimensional data modeling. The results demonstrate that the model has a good generalization effect in the generation of chord music, but some high-quality music clips are rare. Chu et al. [12] proposed a hierarchical neural network model for generating pop music based on note elements. The lower layer handles melody generation, and the upper layer produces chords and drums. Two practical applications of this model are related to neural dance and neural storytelling at the cognitive level. However, the shortcoming of this model also lies in the note-based generation mode, which does not include music theory research, thus limiting its musical creativity and stylistic integrity. Lattner et al. [25] learn the local structure of music by designing a C-RBM architecture that utilizes convolutions only in the temporal dimension in order to model time invariance, instead of pitch invariance, breaking the concept of pitch. Its core idea is to grammatically reduce the structure of music generation before music generation, such as music mode, rhythm pattern, etc. The disadvantage is that the

musical structure is plagiarized. Huang et al. [26] proposed a transformer-based model for music generation. The core of the algorithm is to reduce the intermediate memory requirement to the length of the linear sequence. Finally, it's possible to generate a combination of tiny segment steps for a few minutes and use it in JBS Choir. Despite the experimental comparison of Maestro's two classic public music datasets, the qualitative evaluation was relatively crude.

2.2. Traditional Chinese Music Computing

To our knowledge, there are few available MIDI-based datasets of Chinese folk music. Luo et al. [33] proposed an algorithm to generate genre-specific Chinese folk songs based on auto-encoder. However, the results could only produce simpler fragments and did not perform qualitive analysis of the music from a musical genre perspective. Li et al. [34] presented a combined approach based on Conditional Random Field (CRF) and RBM for classifying traditional Chinese folk songs. Such an approach is noteworthy for being the first in-depth qualitative analysis of the classification results from a music-theoretical perspective. Zheng et al. [35] reconstructed the speed-update formula and proposed a Chinese folk music creation model based on the spatial particle swarm algorithm. Huang [36] collected data from two musical elements based on Chinese melody, and analyzed the application value of Chinese melody imagery in creating traditional Chinese folk music. Zhang et al. [37,38] conducted music data textualization and cluster analysis on Chinese traditional pentatonic groups. In summary, our motivation is to produce Chinese pentatonic music with hierarchical structure and local pentatonic music with multiple musical features and uniform rhythms, as shown in Figure 1 [39].

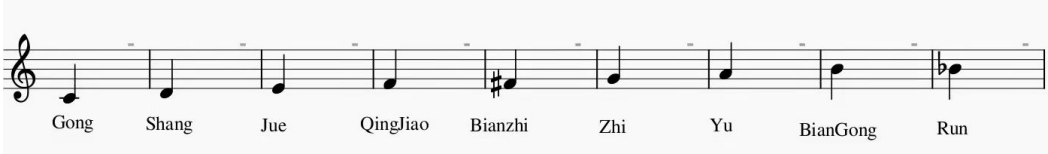


Figure 1. The five main scales and four partial scales in traditional Chinese pentatonic music.

3. Preliminaries

Motivated by generating music in a specific genre, it is crucial that the dataset needs to be addressed. Faced with the problem of stylistic identification of the folk music corpus collected in Section 4, we consider it unacceptable to make hasty assumptions about their identification. With this motivation, we wish to make a preliminary analysis of the genre using a data mining approach, as a support for a specific style music generation task. Afterwards, we will perform our music generation task based on this dataset. This is a reliable way of thinking.

Suppose that, faced with a new problem, we may wish to build a probabilistic model to understand the basic properties of the phenomenon. This is a tentative data analysis process, and clustering algorithms are created for such purposes. While it provides the basis for our overall goal of analyzing musical genres, it remains a tentative and experimental approach. We must explore more details of the information in the context of music theory. Of particular relevance to our work is the fact that the development of clustering algorithms has taken a major step forward with the advent of many types of clustering algorithms. To improve the robustness of our experiments, we selected three types of clustering methods based on division, density, and hierarchy. For the music generation task, we mainly use a compound architecture based on BiLSTM and CNN. Hence, we will launch our related technology description.

3.1. K-Means

Suppose a given data sample X contains n objects with m-dimensional attributes $X = X_1, X_2, X_3, \ldots, X_n$. The K-Means algorithm will cluster K objects into specified clusters based on the similarity between n objects. Each object can only be included in one cluster

with the smallest distance from the cluster center [40]. The K-Means algorithm first needs to initialize K cluster centers C_1, C_2, C_3, ..., C_n, $1 < j \leq K$, and then calculate the Euclidean distance between each object and each cluster center as follows, with a distance of:

$$\mathrm{dis}(X_i, C_j) = \sqrt{\sum_{t=1}^{m} (X_{it} - C_{jt})^2} \tag{1}$$

In the Formula (1), X_i represents the object i , $1 < i \leq$ n, C_j represents the attribute of the cluster center j, $1 < j \leq k$, X_{it} represents the attribute t of the object i, $1 < t \leq m$, C_{jt} represents the attribute t of the cluster center j. Compare the distance of each object to each cluster center in turn, assign the objects to the clusters of the nearest cluster center, and obtain K clusters S_1, S_2, S_3, ..., S_K.

The algorithm uses the center to define the prototype of the cluster. The center of the cluster is the average value of all objects in the cluster in each dimension. The calculation formula is as follows:

$$C_l = \frac{\sum_{X_i \in S_l} X_i}{|S_l|} \tag{2}$$

In the Formula (2), C_l represents the center l of the cluster, $1 < j \leq k$, $|S_l|$ represents the number of objects in clusters l, X_i represents the object i in clusters l, $1 < i \leq |S_l|$.

3.2. OPTICS

OPTICS is a density-based clustering algorithm, which is improved on DBSCAN. Because the DBSCAN algorithm is very sensitive to the initial parameters, the influence of the first two parameters on the calculation results is crucial [41]. Some concepts of the OPTICS algorithm are defined in the same way as DBSCAN, such as σ-Neighborhood, core object, direct density, density reachability, and density connection [42] (assuming our sample set is):

σ-Neighborhood: For $x_i \in X$, its σ-neighborhood contains the sub-sample set whose distance between x_j and in the sample set is not greater than σ. σ-Neighborhood is a set, expressed as follows, the cardinality of this set is recorded as $\left|N_\sigma(x_j)\right|$.

$$N_\sigma(x_j) = \{x_i \in X \mid \text{ distance } (x_i, x_j) \leq \sigma\} \tag{3}$$

Core object: For any sample $x_i \in X$, if its σ-Neighborhood corresponds to $N_\sigma(x_j)$ at least contains MinPts samples, if $\left|N_\sigma(x_j)\right| \geq MinPts$, then x_j is the core object.

Density direct: If x_i is located in the σ-neighborhood of x_j and x_j is the core object, it is said that x_i is to be direct by density x_j. The opposite is not necessarily true; it cannot be said to be directly reached by density at this time unless x_i is also the core object, and direct density does not satisfy symmetry.

Density is reachable: for x_i and x_j, if there is a sample sequence p_1, p_2, ..., p_r that satisfies $p_1 = x_i$, $p_r = x_j$ and p_{t+1} is directly reached by the density p_t, and it is said that the density is reachable. In other words, the density can be reached to meet the transitivity.

Density connected: For x_i and x_j, if there is a core object sample x_k, x_i are reachable by the density x_k; it is said that the density between x_i and x_j are connected. The density connection relationship satisfies symmetry.

On the basis of the above definition of DBSCAN, OPTICS has introduced two definitions required by the algorithm:

Core distance: For a sample $x \in X$, given σ and $MinPts$, the smallest neighborhood radius that makes x a core point is called the core distance of x. Its mathematical expression is as follows, in which $N_\sigma^i(x)$ represents the node i closest to the node x in the set $N_\sigma(x)$, such as $N_\sigma^1(x)$ in $N_\sigma(x)$ and Nearest node x.

$$cd(x) = \begin{cases} undefined |N_\sigma(x)| < MinPts \\ d(x, N_\sigma^{MinPts}(x)) |N_\sigma(x)| \geq MinPts \end{cases} \tag{4}$$

Reachability-distance: Let $x, y \in X$, for a given σ and $MinPts$, the reachability-distance of y about x is defined as:

$$rd(y,x) = \begin{cases} undefined | N_\sigma(x)| < MinPts \\ max\{cd(x), d(x,y)\} | N_\sigma(x)| \geq MinPts \end{cases} \quad (5)$$

In particular, when x is the core point (the corresponding parameters), it can be understood $rd(y,x)$ according to the following Formula (6):

$$rd(y,x) = min\{\eta : y \in N_n(x) \& |N_\eta(x)| \geq MinPts\} \quad (6)$$

The Birch algorithm is a traditional and efficient hierarchical clustering algorithm that does not require pre-setting the entire dataset, due to the fact that the Birch algorithm scans the dataset in an incremental and dynamic manner to construct the clustering feature tree (CF tree) of the dendrogram [43]. Hierarchical coalescence and iterative relocation are combined by it, using a bottom-up hierarchical algorithm, and then the results are updated using iterative relocation.

It has two main stages, including the construction of a memory tree by scanning the dataset and then applying the algorithm to the cluster leaf nodes. The CF tree is a height balanced tree controlled by two parameters, a branching factor B and a threshold T, constructed while scanning the data. When a data point is encountered, the CF tree is traversed by selecting the nearest node at each level starting from the root.

3.3. BiLSTM

Long Short-term Memory is a special Recurrent Neural Network that has been widely applied in natural language processing, machine translation, and music generation [44]. Similar to RNN, it interacts in a chained pattern, but it has a more complex internal structure. A remarkable advantage of the LSTM is that it avoids the problem of RNN being restricted to the existence of simple *tanh* layers inside the structure and simple stacks of external neural units. When RNN solves various sequential problems, the modules undergo extensive gradient explosion and are unable to preserve long-term dependencies, while LSTM provides an efficient way to deal with this problem.

The LSTM has the capability to selectively decide which information is allowed to pass or be prevented from passing (i.e., remember this information or forget this information) by controlling cell states with structures called gates. A layer of sigmoid functions and a point multiplication operation form the basic structure of the gate. where i_t is the input gate, o_t is the output gate, f_t is the forget gate, and C_t is the cell vector. The specific structure is shown in Figure 2.

$$i_t = \sigma(W_x^i\, x_t + W_h^i\, h_{t-1} + W_c^i\, c_{t-1}\,) \quad (7)$$

$$f_t = \sigma(W_x^f\, x_t + W_h^f\, h_{t-1} + W_c^f\, c_{t-1} + b_f\,) \quad (8)$$

$$o_t = \sigma(W_x^o\, x_t + W_h^o\, h_{t-1} + W_c^o\, c_{t-1} + b_c\,) \quad (9)$$

$$c_t = f_t\, c_{t-1} + i_t\, tanh(W_h^c\, h_{t-1} + W_x^c + b_c) \quad (10)$$

$$h_t = o_t\, tanh(c_t) \quad (11)$$

Derived from the computation process of LSTM, Bidirectional Long Short-term Memory (BiLSTM) adds the reverse sequence operation. It can be understood that the input sequence is reversed, and the output is calculated again in the way of LSTM. The final result is a simple stack of the results of the forward LSTM and the reverse LSTM. In this way, the model enables a comprehensive consideration of contextual information, which idea motivates exploring it for potential semantic learning of musical contextual segments.

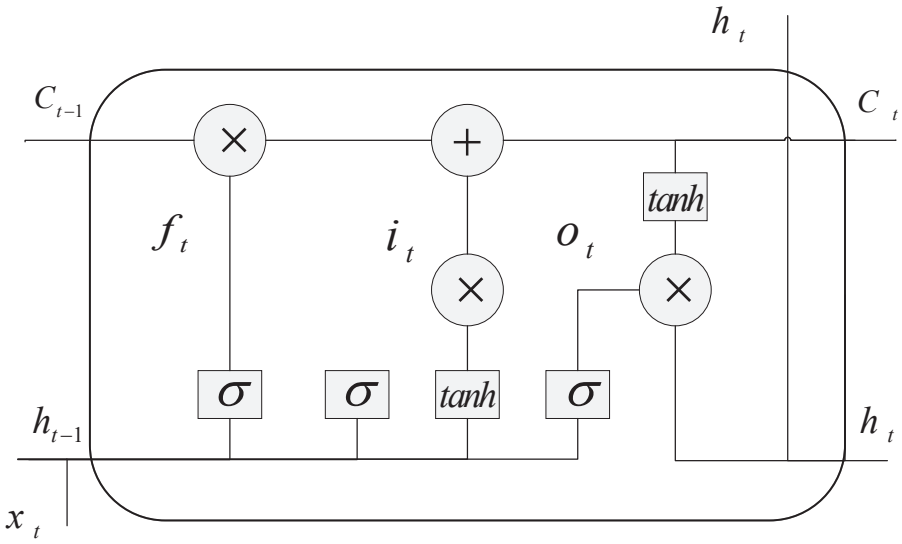


Figure 2. The internal structure of long short-term memory unit.

3.4. CNN

The deep learning method represented by the Convolutional Neural Network realizes object recognition and classification, and the feature extraction is completely handed over to the machine [45]. Feature extraction is achieved through the convolution of different filters, so that distortion and data features can be maintained to a certain degree of invariance, and scale invariance can be achieved through maximum pooling layer sampling. While maintaining the three invariants of traditional feature data, the extraction method minimizes the details of manual design and utilizes the computational power of the computer to actively search for suitable feature data through supervised learning. In the field of image classification, the classification results should be similar no matter where the target features in the image are translated to, which is derived from the Euclidean geometric transformation [20]. However, music has a special quality similar to Euclidean geometric transformations, which is called transposition invariance. Thus, CNN is regarded as an enlightening method to solve the transposition invariance of music [14].

The basic components of the convolutional neural network include an input layer, convolutional layer, pooling layer, activation layer, and fully connected layer, as shown in Figure 3. The convolutional layer extracts features by translating on the original image, the activation layer increases the nonlinear segmentation ability, and the pooling layer compresses the amount of data and parameters, reduces overfitting, and reduces the complexity of the network.

$$h_{ij}^k = (W^k * x)_{ij} + b_k \tag{12}$$

$$f[m,n] * g[m,n] = \sum_{u=-inf}^{} \sum_{v=-inf} f[u,v] * g[m-u, n-v] \tag{13}$$

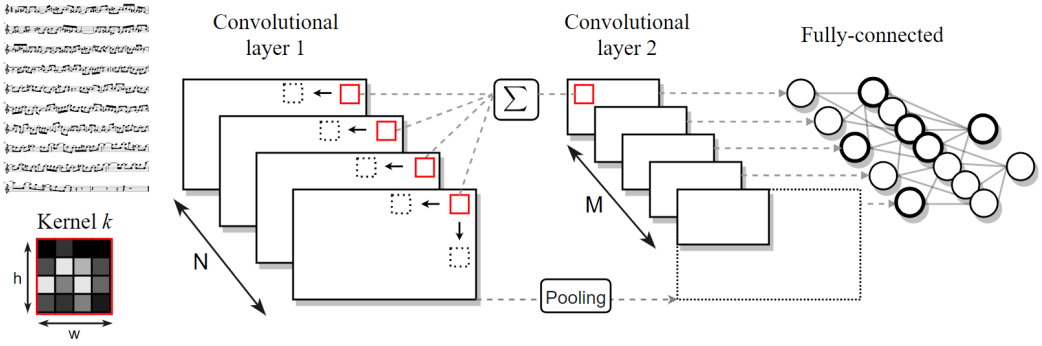


Figure 3. The basic operation process of convolutional neural network.

4. Dataset

Our objective is to generate traditional Chinese folk music based on the properties of deep learning architectures. We must identify the collected music corpus genres; in other words, it must be identified as belonging to traditional Chinese folk music. Therefore, semantic analysis of the music corpus is essential to be carried out as a fundamental work. However, because of the considerable amount of data in the music corpus, clustering algorithms are often preferred for their unsupervised and efficient exploration of data patterns. We expect to conduct clustering experiments with the assistance of traditional Chinese folk music theories, from which we initially explore whether the music corpus we acquired is appropriate. Inspired by text clustering, we advanced a new research idea to transfer the method of text clustering to symbolic music data to achieve our purpose.

4.1. Data Acquisition

As opposed to waveform music, this work is based on the symbolic music. We crawled approximately 1300 pieces of traditional Chinese folk music from the Internet to build an initial corpus. Since all the music pieces were not distinguished carefully when they were crawled, their time signatures and scores were uncertain. They needed an alignment operation to limit their average length to about 12 s. The py-midi and music21 toolkits were used to pre-analyze the data and perform a rough pre-processing of this Chinese traditional folk music dataset, including sequence alignment. Before generating the music, the ethnicity of its text dataset was first analyzed. The MIDI file needed to be converted into a score object to extract the instruments of each track. Then the notes of the piano tracks were extracted and then the pitches of the notes under different scales, i.e., the ranges, were reported. The pitch range of notes is approximately between one to three sets of minor characters. Thus, the folk music dataset was further transformed into a textual music dataset. (Note that in generating the music, we used a pre-processed traditional Chinese folk music dataset in MIDI format, which is the TCPS music dataset we built.)

4.2. Data Pre-Processing

Since this paper first performs text clustering on symbolic music, the current mainstream clustering algorithms mainly operate on structured data; thus, it is necessary to structure the irregular text data. Here, we use the classical weighting method tf-idf for text mining and information retrieval, which represents the product of term frequency (TF) and inverse document frequency (IDF). The specific calculation formula is as follows.

$$tf - idf = tf * idf \tag{14}$$

$$f_{ij} = \frac{n_{ij}}{\sum n_{kj}} \tag{15}$$

Formula (19) shows that $n_{i,j}$ is the number of times the word appears in the document d_j, and the denominator is the sum of the times of all words in the document d_j.

$$idf_i = log \frac{|D|}{j : t_i \in d_j} \tag{16}$$

Formula (20) shows that $|D|$ is the total number of files in the corpus and $\{j : t_i \in d_j\}$ represents the number of files containing words t_i (the number of files with $n_{i,j \neq 0}$).

Through tf-idf calculation, the note elements were weighted. The first step was to extract the complete notes for textualization, which means that they have different scales. After processing, 19 features were extracted, which were distributed on the piano keyboard. In the interval from one to three groups of small characters; then, in order to simplify the data and facilitate clustering statistics, the pitch of its notes was extracted. After vectorizations, the number of features was extracted, and then converted to a weight matrix calculation : $87,000 \times 19$-dimensional tensor.

In the preparation stage of music generation, this paper discards the music of other time signatures when preprocessing the midi data, and only selects the music of 4/4 time. After this step, we cleaned up the selected music and trimmed the longer music clips to the appropriate size, around 12 s. Finally, we encode the segmented piano roll segments into music vector data. Note that in this work, we consider the mathematization of piano roll fragments at the level of musical meta. It is worth noting that while we only use our model to generate fixed-length segments, this model is capable of generating music of any length based on input constraints. Since very low and very high notes are uncommon, we discard notes below C1 or above C8. Therefore, the size of the target output tensor (i.e., the artificial piano roll of the segment) is 4 (music bars) × 96 (time steps) × 84 (notes).

4.3. Clustering Evaluation Metrics

To evaluate the impact of the clustering algorithm on the textual music dataset and to analyze the nationality of the TCPS music dataset, we selected seven objective metrics to evaluate the clustering model both externally and internally. Internal evaluation measured the urgency between sample points of a cluster and the distance between samples and other clusters. External evaluation compared it with a reference model. Figure 4 shows the classification of the metrics.

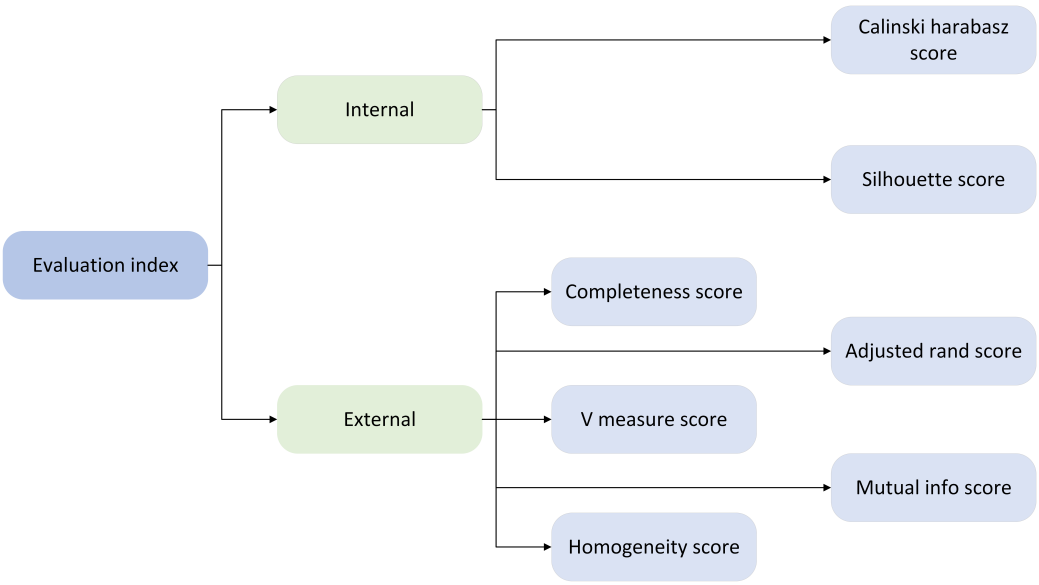


Figure 4. Seven internal or external metrics for evaluation of clustering algorithms.

4.4. Experiment and Results

Given our focus on traditional Chinese folk music, we followed the motivation presented at the beginning Section 4, i.e., to implement the concept of clustering symbolic music in pursuit of identifying the collected musical corpus genres through data mining. First, in the data pre-analysis section, we performed text-weighted pre-processing on the music dataset in TCPS MIDI format. The ethnicity of Chinese traditional folk music was analyzed and verified on this dataset. Then, in order to demonstrate the wide applicability of different clustering models in this dataset and the objectivity of clustering, we used three clustering methods to cluster the music data from three perspectives: division, density, and hierarchy. Among them, the k-value clustering determined by the division-based K-means clustering method achieves the smallest squared error, with dense clustering results and significant inter-class differences. However, the difficulty in choosing k-values or the sensitivity to noisy and isolated points may lead to numerous iterations and long time-consuming periods. Instead of explicitly generating clusters of data, OPTICS simply sorts the objects in the data object set and then computes an ordered list. It contains a lot of information for extracting clusters and is insensitive to the parameter transformations in the

clustering process. However, they are not as good as K-means and Birch as far as clustering results are concerned. There is more computational efficiency in the Birch algorithm, which saves more memory and clusters better than OPTICS, as shown in the Table 1.

Table 1. Comparison of clustering efficiency results.

	K-Means	OPTICS	Birch
Clustering time (sec)	6.5	2.2	2.0

In the Chinese seven-note scale, the scale composition is mainly composed of a pentatonic scale, including *Gong*, *Shang*, *Jue*, *Zhi*, *Yu* , to which two of the other four partials *Qingjiao*, *Bianzhi*, *Biangong*, *Run* are added. The pitch name system is expressed as *C*, *D*, *E*, *G*, *A*, *F*, *F#*, *B*, *Bb* [46].

There are three sources of Yayue in all dynasties. It was inherited from the court music works of the Zhou dynasty, reconstructed according to the music theory of the Zhou dynasty, or newly produced on the basis of new vocal and popular music. It is made up of the following elements: *Gong*, *Shang*, *Jue*, *Bianzhi*, *Zhi*, *Yu*,, and *Biangong* [47]. Qingyue is also known as Qingshang music. It is a traditional music that emerged in the Three Kingdoms, Jin Dynasty, Southern and Northern Dynasties and dominated the musical life of the time. It consists of *Gong*, *Shang*, *Jue*, *Bianzhi*, *zhi*, *Yu*,, and *Biangong* [48]. Yanyue was a very artistic song and dance music that provided entertainment and appreciation during the palace banquets from the Sui, Tang and Song dynasties. The court Yanyue of Sui and Tang dynasties reflects the highest achievement of music culture in this period. It comes from the continuous accumulation of traditional music of the Han nationality and the large-scale input of foreign music since the Han and Wei Dynasties. It is composed of *Gong*, *Shang*, *Jue*, *Bianzhi*, *Zhi*, *Yu*, and *Biangong* [49].

We back-track the data to trace the components after clustering, and we go back to the first seven musical elements in each cluster, which correspond to the five main tones and four partial tones of traditional Chinese national music [50]. The results revealed that cluster 1 possessed the structural characteristics of the Yanyue mode, clusters 2, 3, 4, 8, and 9 has the structural characteristics of the Qingyue mode, and clusters 5, 6, and 7 features the structural characteristics of the Yayue mode, as shown in the Table 2.

Table 2. The main note in the cluster and other notes.

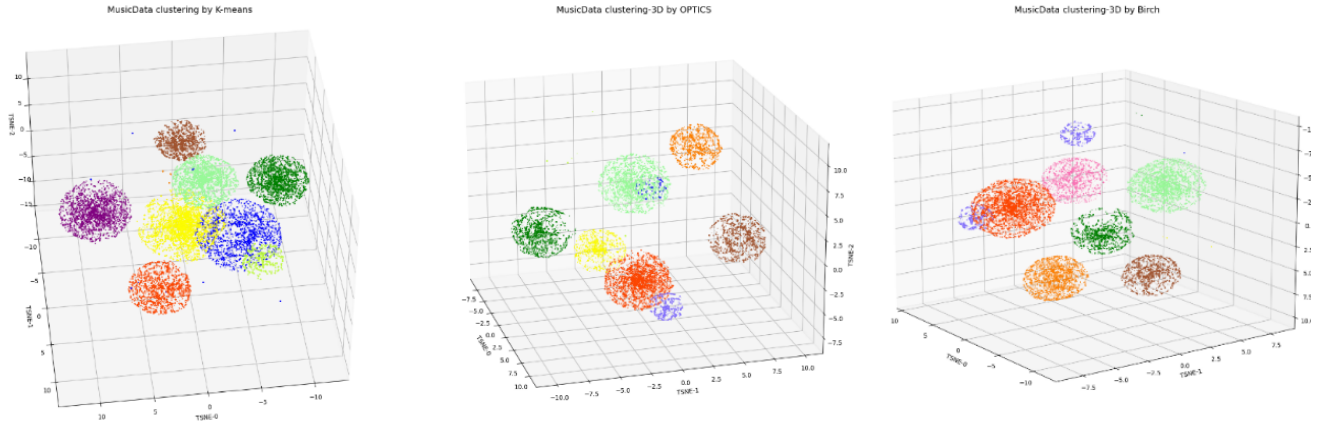
	Main Note	1	2	3	4	5	6
Cluster 1	Bb	B	E	G	F	D	C
Cluster 2	A	G	F	E	D	C	B
Cluster 3	D	G	F	E	C	B	A
Cluster 4	E	G	F	D	C	B	A
Cluster 5	G	F#	F	E	D	D	B
Cluster 6	F#	C	G	F	E	D	B
Cluster 7	F	F#	G	E	D	C	B
Cluster 8	B	G	F	E	D	C	A
Cluster 9	C	G	F	E	D	B	A

Figure 5 shows the three-dimensional visualization results of clustering under K-Means, OPTICS, and Birch. It can be observed that the figure matches the objective evaluation metrics in Table 3. k-Means and OPTICS demonstrate relatively better aggregation results. Similar results indicate that annotated text data are more suitable for partition and density-based clustering methods.

Table 3. Comparison of objective evaluation index results of different clustering methods.

	Sil	Cal	Hom	Com	V_me	Adj	Mut
K-means	0.930	293,355.316	0.934	1.000	0.966	0.949	0.966
OPTICS	0.980	165,675.291	0.949	1.000	0.974	0.972	0.974
Birch	0.861	158,523.624	0.859	1.000	0.924	0.878	0.924

It is evidenced by the clustering experiments that the symbolic music dataset we collected belongs to the traditional Chinese pentatonic scale music, so we call it TCPS dataset. This dataset will later be used for the traditional Chinese pentatonic scale music generation task.



(**a**) (**b**) (**c**)

Figure 5. 3D clustering visualization results performed by K-Means (**a**), OPTICS (**b**), and Birch (**c**). The T-Distributed Stochastic Neighbor Embedding (T-SNE) visualization method is applied to the visualization of three methods.

5. Proposed Model

5.1. 3D-SCN

A three-dimensional sequential convolutional network for predicting the output of a joint time-series distribution is described by combining the ideas of a time-series model and an image convolutional model in a music generation model. Since music has musical properties, such as temporal fluidity and tonality (when harmonies and melodies in one key transition to harmonies in another key, i.e., interval adjustment, and the progression is kept at a certain length) and intensity, we can call it a transition. It is not an absolute control, so the relative position of a group of coherent harmonies and the relationship between notes is crucial, as in C major to D major. It is only possible to predict notes but not to explore the potential relative relationships between chords using a simple time series model. Instead, it is possible to use a convolutional neural network for learning the relative relationships between core musical semantics by feeding a set of note data, either convolutionally or cross-correlationally. The relationship between the two-dimensional sequences i and j can be represented as follows:

$$(i * j)_n = \sum_{m=-\infty}^{+\infty} i_m j_{m+n} \tag{17}$$

If input i or j is offset, $(i * j)_n$ will also be offset by σ according to the translation invariance in convolutional neural networks. Let us use this idea for music generation via translation invariance. The instantiated notes in the model can be designed to be related to each other, assuming there is a note vector v_t at time t and a note vector v_{t+1} at the

next time $t+1$. Through translation invariance, the sum of vectors w_t and w_{t+1} has the relationship of $w_t = v_t^{(i+\sigma)}$ and $w_{t+1} = v_t^{(i+\sigma)}$, and the output model satisfies:

$$p(\hat{w}_{t+1}|w_t) = p(\hat{v}_{t+1}|v_t) \tag{18}$$

In order to achieve the translation invariance of the note sequence and perform the joint distribution output of the note sequence, the three-dimensional sequential convolutional neural network (3D-SCN) described in this paper, which is constructed in three dimensions, is based on BiLSTM. The 3D training flows from three directions, as shown in Figure 6, performing single-note binarization encoding input for each network unit. The input format is a fifteen-bit binary code from the note axis. The first twelve bits represent an octave, and each bit represents its twelve semitones, with 1 for playing a note and 0 for not playing a note. The last two digits 00 indicate that when the current last note is the same in a half measure, play two quarter notes intermittently, and the last two digits 01 indicate that when the current last note is the second note, continue playing two quarter notes. Same in the 1/2 measurement. A quarter note, the last two numbers of 10, means that a half note can be played directly when the current last note is the same in a half measure. The last digit 0 means arpeggiated fingering is not active, and 1 means arpeggiated fingering is active. In order to mine the relationship between notes, we use a partial window $d^{(n,f)}$ as input in the model, which contains the partial note vector $v_t = [v_1, v_2, \ldots, v_n]$, that is, when $1 < i \leq 25$, $d_i^{(n,f)} = v_{n-13+i}^{(t)}$, if at time t , $d^{(n,f)}$ exceeds the bounds of $v^{(t)}$. That is, no notes are played. The window is out of bounds. The out-of-bounds value will be set to 0.

A sliding interval window time t contains an array window $s^{(n,t)}$, consisting of notes two octaves before and after, indicating that all octaves are offset by i on the note count played at time t, that is:

$$S_i^{(n,t)} = \sum_{m=-\infty}^{+\infty} v_{i+n+12m}^{(t)} \tag{19}$$

With the two initial input directions of the 3D-SCN, the first propagates from top to bottom along the time axis, while the next time step propagates along the note axis. Special attention should be paid to the use of BiLSTM, which performs cell input training and allows feature extraction along the forward and reverse input directions perpendicular to the third direction of note and time. The note feature data during the learning process also has information about the past and the future, because music and language are similar in this respect, and whether the harmony between different chords implies certain rules. The model draws on the analogy between language modeling and music modeling, and combines the feature vectors of two notes in positive and negative order into the final feature of the expression $v_i^{(t)} = [{}^{r(t)}_{v_{ij}} \oplus {}^{s(t)}_{v_{ij}}]$. The input of the previous step received by the note axis is the output of the previous note layer after activating the note on the time axis and conducting the previous note calculation in the previous BiLSTM. The final activation function of the note axis is to use the softmax function to obtain the probability $p^{(n,t)}(v_n = 1|v < n)$. Each time step has a bidirectional LSTM note axis network with corresponding bound weights. Musical semantics are the forward and reverse modeling of upper and lower notes in a single time step. While jointly distributed in space, each note has a separate temporal network with bound weights to model musical temporal relationships. The specific architecture of 3D-SCN for 3D spatiotemporal modeling is shown in Figure 6 [51].

5.2. BoYaTCN

CNNs provide an efficient way of recognizing complex objects, such as pedestrian recognition, by stacking layers to abstract object features. In practice, however, the norm of the error gradient decreases with each layer until it reaches a point so small that it no longer allows optimizing the network parameters, specifying a maximum threshold for creating depth in classical CNNs. To alleviate this problem, residual networks were proposed. It

relies on a residual connection between the convolutional layers, which adds the output of the previous layer to the output of the stacked layers. Thus, the output of a layer in the residual network can be formally defined as:

$$y = F(x, \{W_i\}) + W_s x \tag{20}$$

where x and y denote the input and output of the layer, F is the residual mapping to be learned, and W is a linear projection for matching the dimensions of x and y when the number of channels differs from one layer to another. While all layers are interconnected from the L layer of the architecture to the $\frac{L(L+1)}{2}$ layer, this will increase the number of direct connections between the layers, thus maximizing the information flow between them; the feature map size is made mandatory for matching. This model has been demonstrated to promote the reuse of features learned throughout the layers, thus reducing redundancy in the information stream.

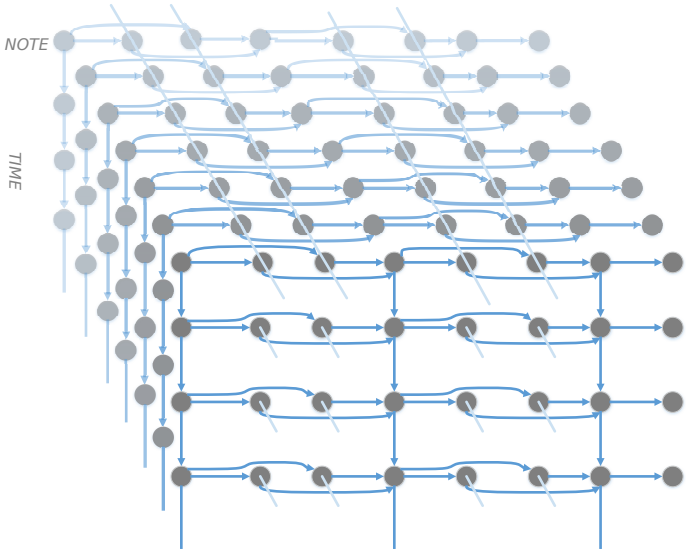


Figure 6. The architecture of 3D-SCN: three-dimensional sequential convolutional network.

While processing sequential information, people learn to unconsciously select the most relevant information from this process. For example, if the given musical fragment's score is read through to find its key, it will focus on specific notes and temporary notations and not on the others. In a broader sense, one is able to assess the interdependence between input data and desired output, which resulted in the adoption of the musical attention mechanism.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{21}$$

where d_k is the dimension of the key and the query, and the main improvement of the algorithm appears when processing larger values of d_k. In fact, the scale factor $\frac{1}{d_k}$ prevents the dot product from growing excessively in magnitude, which causes the gradient of the softmax function to vanish. Multi-head attention overcomes this problem by going further than performing a single attention function using the model dimensions K, V, and Q. Instead, they rely on different linear projections' learning. The d_k dimension is used to query the values of the key and d_k dimensions. The attention functions are then applied in parallel to all projection versions of K, V, and Q to output values in the d_k dimension. Finally, they are concatenated and linearly projected to obtain the final output.

$$Multi - Head(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{22}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{23}$$

$$W_i^Q \in \mathbb{R}^{d_{model}} \times d_Q \tag{24}$$

$$W_i^K \in \mathbb{R}^{d_{model}} \times d_K \tag{25}$$

$$W_i^V \in \mathbb{R}^{d_{model}} \times d_V \tag{26}$$

$$W_i^O \in \mathbb{R}^{hd_V} \times d_{model} \tag{27}$$

where W_i^Q, W_i^K, V_i^V, and W_i^O are the projections of Q, K, V, and output, respectively.

A chord can be composed of several notes while the root note is different, yet the same chord can be played; it is just the relative scales of the notes that are different, which leads to the nature of musical transposition.

The notes do not fundamentally change the relative properties of the music, hence the need to consider transposition invariance in the process of music coding, making note coding flexible and versatile. While in the background of polyphonic music, it is also possible to classify each element vertically according to its importance within the same piano roll framework. Therefore, attention mechanisms specifically embodied in the selective adjustment of musical sequence data from the temporal aspect and the chordal aspect are proposed.

Several critical aspects differ from the multi-headed attention mechanism. It is applied separately at each layer to select the most relevant information. However, this information is not provided to the subsequent layers, but mixed and processed separately again, and the resulting output is used to modulate the last fully connected layer of the network. Each feature map is processed separately by a linear transformation that reduces its dimensionality to d_t, resulting in a group of keys K and values V , and the output of the preceding convolutional layer defines the query Q. The network computes the scaled dot product attention for each reduced feature map as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_t}})V \tag{28}$$

Finally, with the purpose of obtaining a prominent message for each given layer, this set of attention vectors is normalized and blended across the kernel. The whole process is illustrated in Figure 7.

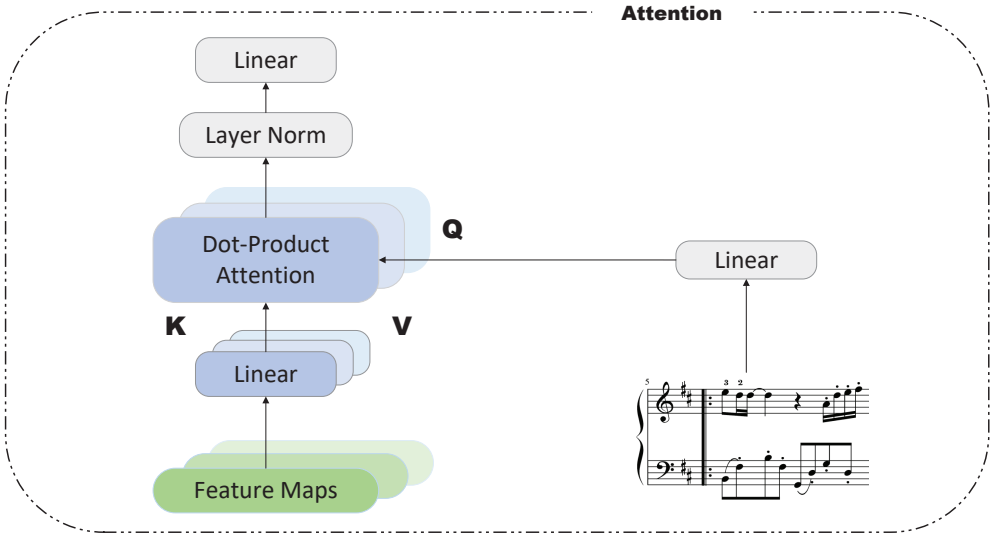


Figure 7. Schematic diagram of the musical attention mechanism.

With the aim of learning the spatial embedding structure of musical attributes, a spatial embedding model based on an encoding network architecture is proposed, which encodes each event in a sequence into an embedding vector and tries to characterize musical events based on it. This is where the embedding architecture is based on a residual convolutional neural network, and a multilayer attention mechanism is introduced for handling transposition features. The point is that the attention mechanism participates in the processing of each convolutional layer separately to obtain a multiscale layered view at different levels of abstraction. The attention module itself operates based on the kernel by processing each feature map independently instead of using spatial attention, and the input is used as a query for all attention modules instead of the feature map itself. Information from different attention levels is concatenated and blended, and then the simpler dot product attention mechanism is used again, as shown in Figure 8. On this basis, a new compound architecture was constructed, named bidirectional octave your attention temporal convolutional network (BoYaTCN), as shown in the Figure 9.

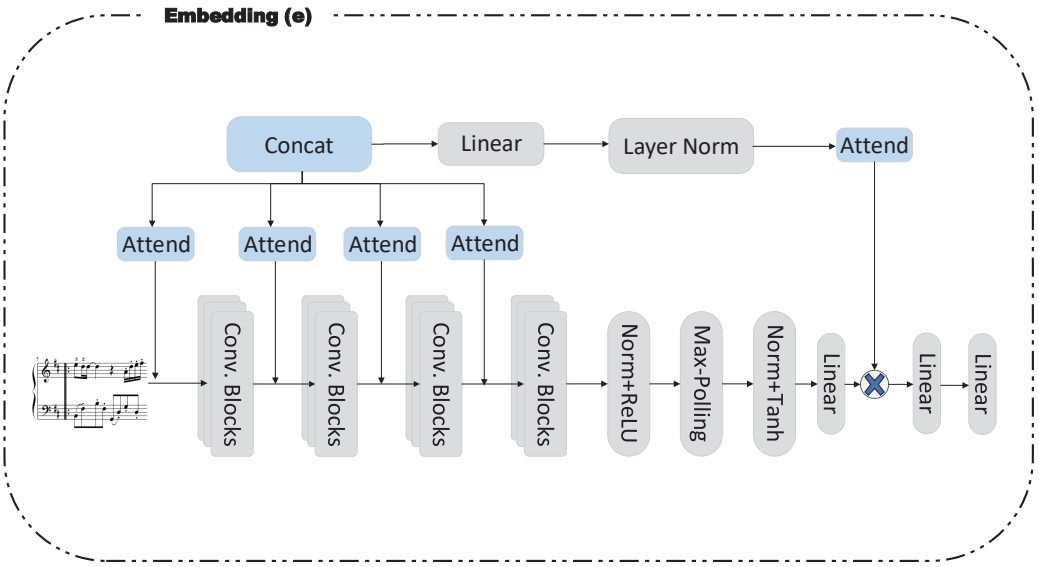


Figure 8. A design of spatial convolutional embedding module for three-dimensional sequential convolutional network.

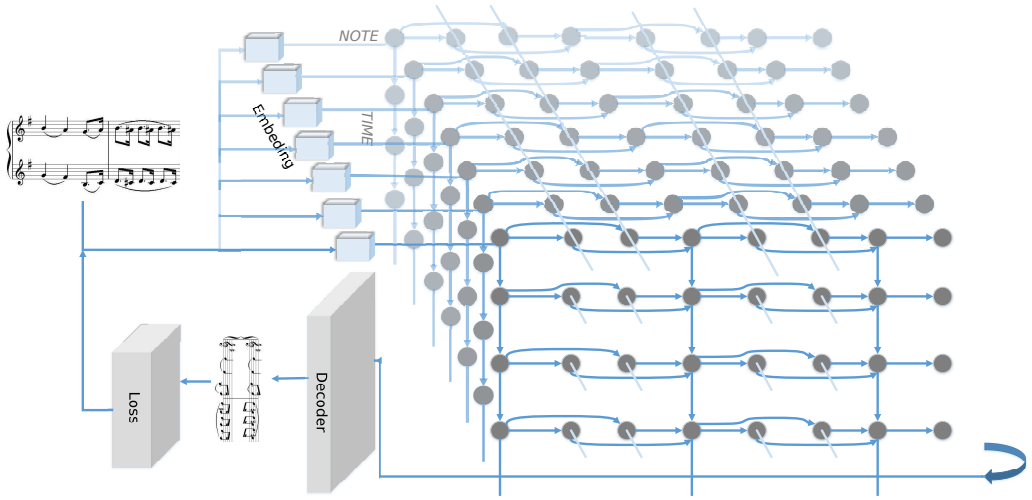


Figure 9. The architecture of BoYaTCN: bidirectional octave your attention temporal convolutional network.

6. Implementation

6.1. Data Representation

In order to simulate the traditional Chinese pentatonic folk music with the coherent characteristics of musical styles, we propose to use the piano roll for representation. The piano roll of a melody is a vector of binary values, and the piano roll with chords can be similar to a score table matrix, in which different pitches can be recorded at different note durations. Formally, an multi-track piano roll of a measure is represented as a tensor $x \in \{0,1\}^{R*S}$, where R and S represent the time step and the number of candidate notes in a measure, respectively. The subsection T piano roll for a measure is denoted as $\vec{X} = \{\vec{X}^t\}_{t=1}^{T}$, where $\vec{X} \in \{0,1\}^{R*S}$ represents the subsection t piano roll distribution for the measure, as shown in Figure 10.

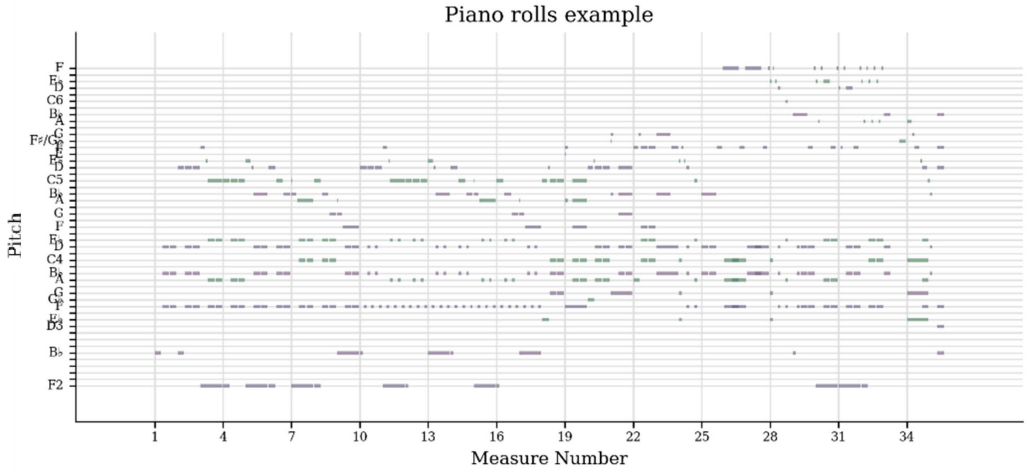


Figure 10. This is an example of a piano roll.

Since the time step of each subsection is composed of smaller time units, the whole large matrix is fixed, which makes the computation of parallel BiLSTM and CNN easier. Figure 11 illustrates an example of music data encoding.

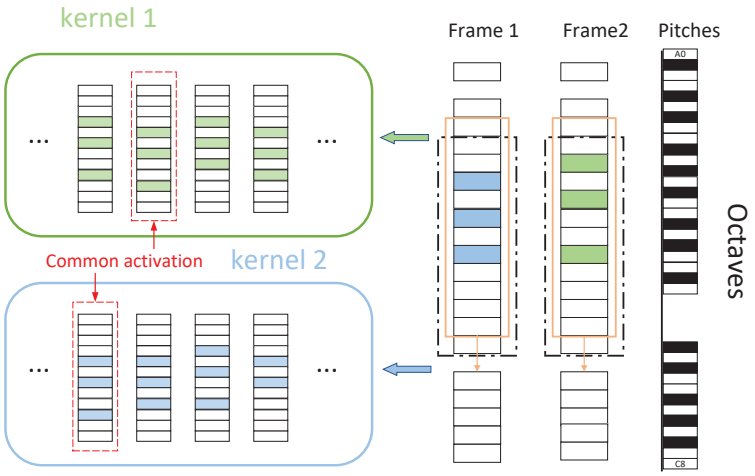


Figure 11. The encoding representation of music data. Piano rolls are transformed into multi-hot vectors for music data embedding operations.

6.2. Model Setting

The experiments were trained using an analytically validated traditional Chinese pentatonic scale music dataset (TCPS), extracting traditional Chinese folk music in 4/4 time, converting it to the key of C, and extracting a piano roll matrix from a MIDI file

by quantifying the duration of 16 frames per beat, setting the played note value to 1 and otherwise to 0. Consequently, the resulting matrix consists of mostly repeated frames. In this frame-based context, a complete model that simply repeats the last music frame outperforms all previously proposed models. To avoid this problem, the whole sequence needs to be predited instead of individual notes. However, the extant models are all evaluated on single-frame prediction tasks, and learning for the entire sequence may make the learning very unstable and invalidate the embedding results of its metric properties. Therefore, the experiments constrain the predictions by converting the piano roll-up matrix to an event-level representation (keeping a single frame for each new event), while following a 60% training dataset and 40% test dataset split.

A 3D-SCN was used for traditional Chinese folk music generation using two BiLSTM layers with 200 nodes each in the temporal direction of temporal relationship modeling. Two BiLSTM layers with 100 nodes each are used in the note direction of spatial modeling. A comparison was made with the TP-LSTM-NADE model proposed by Daniel [14]. Further, the training test was performed using the analytically validated Chinese traditional folk music dataset TCPS, which extracts traditional Chinese folk music in 4/4 time, converts it into C major or C minor, randomly selects 60% for the training dataset and 40% for the test dataset, the dropout is set to 0.5, the maximum paradigm weight constraint is 4, and the optimization is performed using the newer Radam optimizer with the initial learning rate is set to 0.002, and the training time is 20 h.

With the event-based order, the experiments use a three-dimensional sequential convolutional network to predict the next musical event. Since the key to the computation is the pre-embedding operation of the music events, the primary goal of the network construction is to consider improving the structure of the embedding to enhance its prediction. The experiments provide the network with small batches of size 32, each element being a sequence of 12 events (128-dimensional frames). All events are encoded in the de-dimensional space, and the 3D sequence network attempts to predict the 13th event of the sequence by relying on a previously set vector of 12 embeddings in one octave. The prediction is then decoded by reverse processing the encoding and the 128-dimensional resultant vector is compared to the true target by BCE loss. The experiments use the Radam optimization algorithm with a dropout set to 0.5, an initial learning rate set to 0.002, and an optimization scheduler that decays the learning rate to half of its original size after 25 rounds of training with a maximum paradigm weight constraint of 4 and a training time of 27 h.

7. Objective Metrics for Evaluation

In order to objectively evaluate the generation of traditional Chinese pentatonic folk music, traditional machine learning evaluation metrics were used in this quantitative study to evaluate the performance of the model, which are accuracy, precision, recall, and binary cross-entropy loss.

Accuracy: It refers to the proportion of correctly classified samples to the total number of samples.

Precision: It refers to the proportion of samples with correct subclassifications out of the number of samples judged as positive by the classifier.

Recall: It refers to the ratio of the number of correctly classified positive samples to the number of true positive samples.

Binary cross-entropy loss: It refers to a loss function often used in binary classification models; the method is usually used to estimate the probability of rare events.

8. Experiment & Results

To evaluate the performance of different parts of the model, several variants of the model were trained by ablation experiments. First, the experiments present a baseline CNN model (with the same architecture as the spatial embedding module proposed in this paper), as well as a model incorporating a residual mechanism (Residual) and a model with a dense mechanism (Dense). Afterwards, different variants of the attention

mechanism in this paper are experimentally evaluated. From changing the type of attention module used using a simple dot product attention mechanism (DP-SCN) and a multi-head attention mechanism (MH-SCN), to using a model that eventually uses the hierarchical attention mechanism based on the one proposed in this paper (BoYaTCN). Eventually, the experiments are performed by simply adding the model (BoYa+TCN) replacing the modulation aspect of hierarchical attention. In addition to the convolutional baseline model (CNN), dot product attention generation model (DP-SCN), and multi-headed attention generation model (MH-SCN), the proposed three1 dimensional sequential convolutional network (3D-SCN) and TP-LSTM-NADE models are compared with the above models in this paper.

8.1. Quantitative Studies

First of all, it can be noticed that the results obtained using the attention and residual modules outperform those obtained through the simple CNN version, but only resulting in a slight degree of accuracy improvement. Assume here that the visual features of the piano roll frames are relatively simple compared to the visual features of the images, spanning fewer feature layers. Therefore, the presence of different connections in the residual module may hinder the learning of the model rather than improve its performance. On the other hand, the different attention modules greatly improve the prediction accuracy of the model. As expected, the results demonstrate that the accuracy of all datasets has been significantly improved even using the simple dot product attention module, although the use of the multi-headed attention mechanism also further improves the performance of the experimentally designed model to a high accuracy, but does not surpass the 3D-SCN and TP-LSTM-NADE models, as shown in Table 4.

The three-dimensional sequential convolution model BoYaTCN and its simplified version BoYa+TCN in this paper outperform all previous architectures, and a stronger modulated signal (relying on the product rather than a simple sum) on the fully connected layer of the attention matrix yields higher accuracy. Figure 12 depicts the testing procedure for the cutting-edge models BoYaTCN and the other three modals, including BoYa+TCN, 3D-SCN, and TP-LSTM-NADE. The above confirms the merits of the present model approach and demonstrates the benefits of building spatial embedding modules and attention mechanisms between elements of a music sequence for efficient encoding and prediction of music data in a low-dimensional space.

Table 4. Comparison of objective machine learning metrics for ablation experiments based on three-dimensional sequential convolutional networks.

Model	Accuracy	Loss	Precision	Recall
TP-LSTM-NADE	99.01	0.782	0.452	0.236
3D-SCN	99.04	0.775	0.422	0.230
DP-TCN	88.72	0.845	0.332	0.381
MH-TCN	89.23	0.824	0.332	0.371
BoYa+TCN	99.08	0.724	0.473	0.183
BoYaTCN	**99.12**	**0.708**	**0.499**	**0.412**

8.2. Qualitative Studies

Figure 13 shows the traditional Chinese folk pentatonic music randomly generated by 3D-SCN and BoYaTCN models, in which the traditional Chinese pentatonic scale *Gong* (red), *Shang* (orange), *Jue* (yellow), *Zhi* (blue), and *Yu* (purple) notes account for about 80%, and the rest are mostly the other four partials of the traditional Chinese folk pentatonic scale: *Qingjiao*, *Bianzhi*, *Biangong*, and *Run*. Among them, it is not difficult to find that the generated traditional pentatonic folk music is characterized by timbral variations. Such a result may be the effect of the translation invariant architecture training of the convolutional network used in 3D-SCN, and the whole story is harmonious. We can call it a transposition

invariance of musical harmony, which can be related to the powerful ability of BiLSTM to understand the semantics of contextual music. In short, this has achieved our desired goal.

Additionally, each of the five tones of the pentatonic scale is composed of tonal patterns. The interval relationships of the pentatonic modes are shown in Table 5.

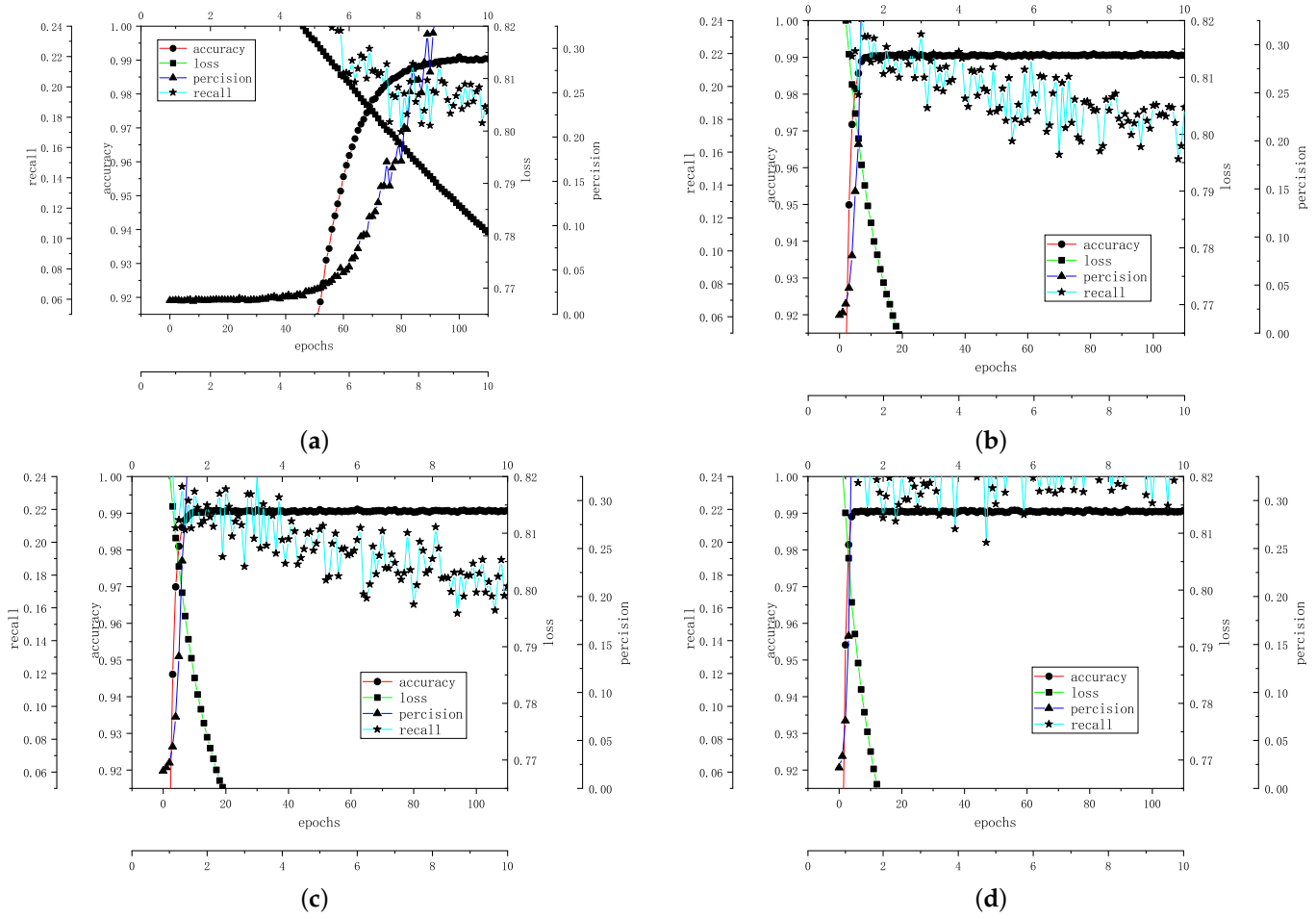


Figure 12. The test dataset evaluation indicators curve of the TP-LSTM-NADE (**a**), 3D-SCN (**b**), BoYa+SCN (**c**), and BoYaTCN (**d**) architecture.

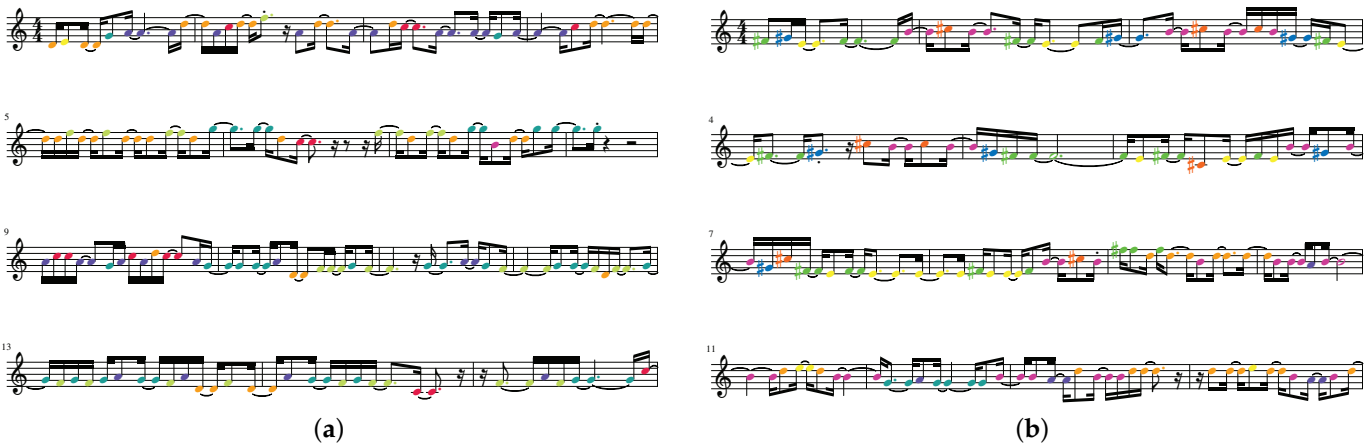
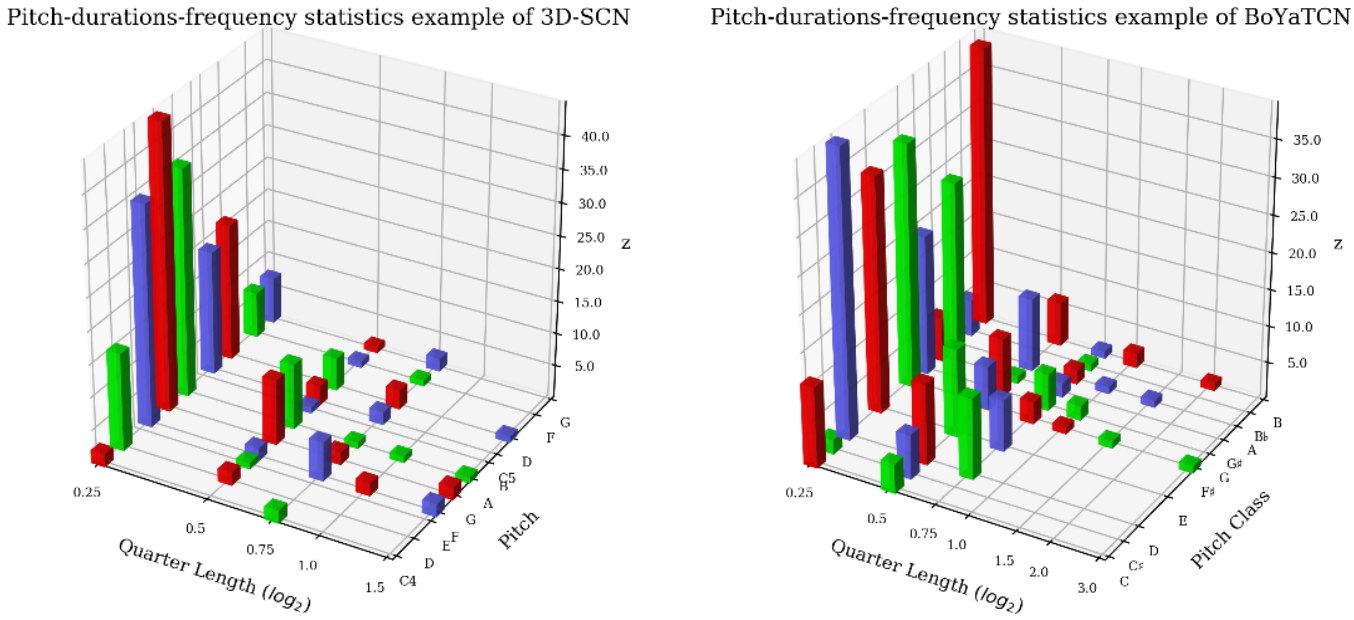


Figure 13. Traditional Chinese Pentatonic scale music generation fragments of 3D-SCN (**a**) and BoYaTCN (**b**).

Table 5. The interval relationship of the traditional Chinese pentatonic scales.

Traditional Chinese KeyScale/Interval	Begin	Second	Third	Fourth	Last
Gong KeyScale	Gong	Shang	Jue	Zhi	Yu
Interval	Major second	Major second	Minor third	Major second	Minor third
Shang KeyScale	Shang	Jue	Zhi	Yu	Gong
Interval	Major second	Minor third	Major second	Minor third	Major second
Jue KeyScale	Jue	Zhi	Yu	Gong	Shang
Interval	Minor third	Major second	Minor third	Major second	Major second
Zhi KeyScale	Zhi	Yu	Gong	Shang	Jue
Interval	Major second	Minor third	Major second	Major second	Minor third
Yu KeyScale	Yu	Gong	Shang	Jue	Zhi
Interval	Minor third	Major second	Major second	Minor third	Major second

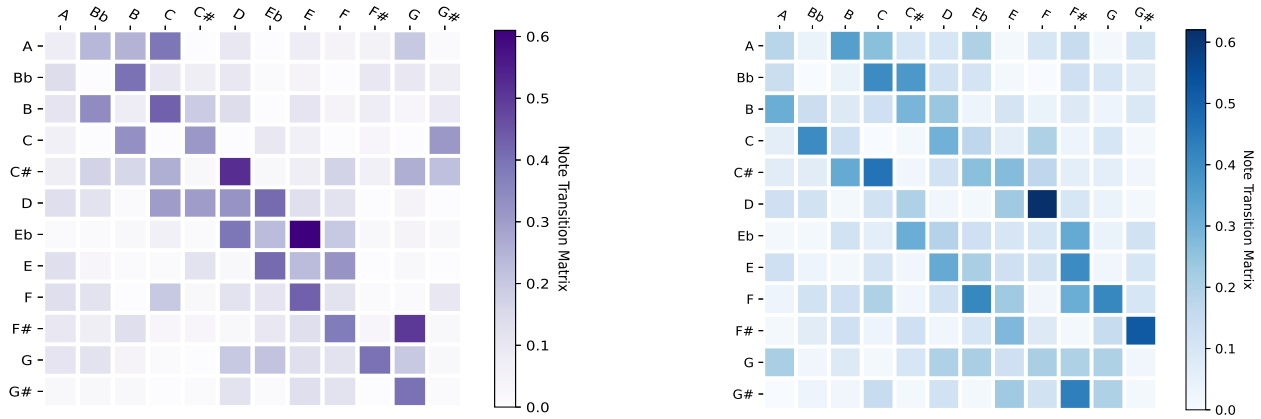
The pitch-duration-frequency statistics graphs of traditional Chinese folk music fragments generated by 3D-SCN and BoYaTCN sequence convolution architectures also illustrate some things. The generated note duration of quarter notes, eighth notes, and sixteenth notes are observed. Looking at the pitch distribution from traditional Chinese pentatonic folk music, they all also appear as *Gong*, *Shang*, *Jue*, *Zhi*, *Yu*, *Qingjiao*, *Bianzhi*, *Biangong*, and *Run*, with the scale composition of traditional Chinese folk music of Qingyue, Yanyue, and Yayue, as shown in Figure 14.



(**a**) (**b**)

Figure 14. Pitch-duration-frequency statistics example of 3D-SCN (**a**) and BoYaTCN (**b**).

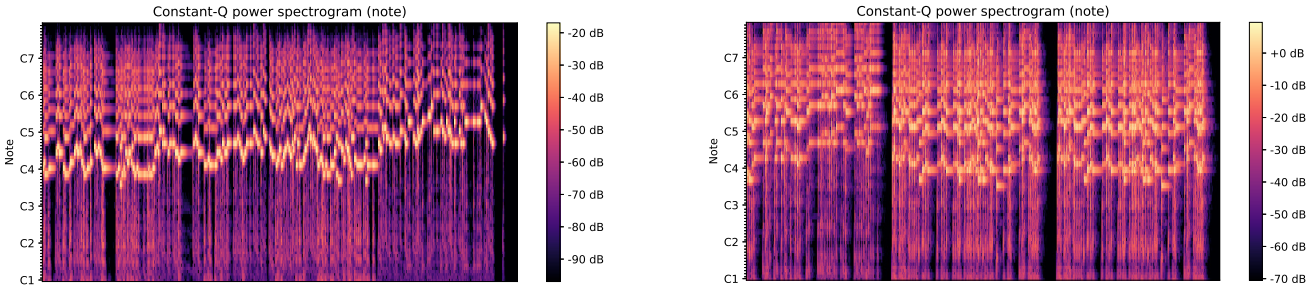
As the style characteristics of Chinese traditional folk music, the interval compositions of the major second and minor third also provide strong support for the experimental generation of the interval composition characteristics of Chinese traditional pentatonic folk music. The note probability transfer matrix of the generated music fragments from 3D-SCN and BoYaTCN shows that the generated music fragments have a major second and minor third as the main interval composition, as shown in Figure 15.



(**a**) (**b**)

Figure 15. Note transition matrix of 3D-SCN (**a**) and BoYaTCN (**b**). The vertical axis indicates the previous note and the horizontal axis indicates the note that follows.

The musical melodies generated by the 3D-SCN and BoYaTCN music generation models were also relatively smooth and stable with even energy distribution. The intervals span across three and a half octaves and range from one-lined octave to three-lined octave, as shown in Figure 16.



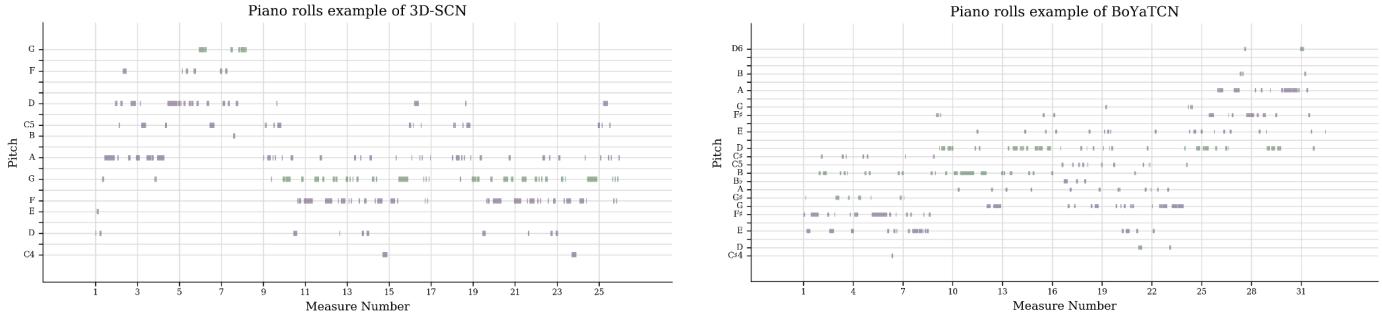
(**a**) (**b**)

Figure 16. Constant-Q power spectrogram of 3D-SCN (**a**) and BoYaTCN (**b**).

The piano roll visualization demonstrated the duration and pitch information of the notes and the distribution composition of the musical phrases, with the vertical axis representing the pitch and the horizontal axis representing the time span of the event. It can be observed that the note-density is relatively reasonable, and the phrases are smoothly articulated and well-structured as the period changes. The phrase bars 4–10 and 10–16 also produce the transpositional quality of the music composition, reflecting the success of the panning invariance of the network architecture design. The structure between measures 4–10 and 10–16 also generates a circular and reciprocal form of musical composition. However, the fragmented notes are of short duration and there are sharp pauses between notes, but the transposition is more evident in phrase bars 19–25 and 25–31. Compared to 3D-SCN, BoYaTCN achieves better results in the structural learning of the music at the phrase level, as shown in Figure 17.

Subjective user research was also conducted in this work. A total of 403 subjects were recruited using Web-based questionnaires requesting professors and students from university and evaluate traditional Chinese pentatonic folk music randomly generated by the proposed model. Of the initial cohort of 403 subjects, 86 were professionals and 317 non-professionals in the music industry. The evaluation questionnaire aims to apply statistics and analyze the five subjective indicators of the generated music: the style integrity, the harmony of the section, whether it has a unified rhythm, whether the section structure is reasonable, and the ethnicity. The data demonstrates that in terms of style integrity, music harmony, and nationality, professionals engaged in music work prefer to generate samples,

while in terms of rhythm and music grammar, the professional evaluation is obviously lower than that of non-professionals, indicating that the model needs to be improved in these two directions. The percentage of people with positive answers is shown in Table 6.



(**a**) (**b**)

Figure 17. Fragment of a piano roll of a random pentatonic scale music sample from 3D-SCN (**a**) and BoYaTCN (**b**).

Table 6. Evaluation of randomly generated music by music professionals and non-professionals.

Model	User Type	Overall Style	Harmonious	Rhythmic	Structured	Nationality
TP-LSTM-NADE	Pro	89.86	93.32	79.57	87.23	85.45
	Non-pro	82.24	90.41	75.63	83.61	82.25
	All	84.42	90.78	76.82	84.68	83.57
3D-SCN	Pro	90.70	96.51	**81.40**	88.37	89.53
	Non-pro	83.60	**95.59**	**84.54**	89.28	86.12
	All	85.11	**95.78**	**83.87**	89.09	86.85
DP-TCN	Pro	77.98	82.56	70.95	78.37	76.41
	Non-pro	73.32	78.72	68.03	73.48	72.26
	All	75.46	80.17	68.49	75.62	74.57
MH-TCN	Pro	79.26	83.24	73.74	80.45	78.38
	Non-pro	74.84	79.33	70.34	75.89	75.56
	All	76.21	80.72	71.98	77.95	76.24
BoYa+TCN	Pro	80.15	85.22	75.41	82.68	82.12
	Non-pro	72.74	82.04	71.62	79.57	80.46
	All	78.71	83.48	72.71	80.06	80.98
BoYaTCN	Pro	**91.21**	**96.67**	82.61	**92.41**	**90.15**
	Non-pro	**84.23**	94.78	85.17	**90.15**	**89.38**
	All	**86.26**	95.29	84.03	**91.34**	**89.74**

9. Conclusions

In this paper, we presented and studied a music generation method based on LSTM, CNN, and musical attention to improve the content generation performance of Chinese folk music. This resulted in a multi-generational selection, and we presented two successful models to implement this idea: three-dimensional sequential convolutional network (3D-SCN) and bidirectional octave your attention temporal convolutional network (BoYaTCN). The main ideas include modeling the note data based on the 3D spatial perspective of BiLSTM in an attempt to learn the temporal embedding of musical events in the context of integrated musical fragments. It also incorporates a multi-layer attention mechanism for designing convolutional modules for music data embedding, both of which capture the relevant musical features for music specificity. We have tested the proposed ablation experiments with multiple parameters on a traditional Chinese folk music dataset. Objective and subjective studies demonstrate that the proposed model is capable of producing pentatonic music that is more melodic and harmonically coherent, and has distinctive traditional Chinese pentatonic tonal characteristics. It also conforms to certain musical syntactic features.

In the future, we plan to expand this research and provide multimode (audio signals, musical notation) musical compositions describing the same semantic content to explore more aesthetically and culturally valuable music generation content. Furthermore, the proposed generative model, with the support of diverse music theories, can also be safely transferred to generative tasks in other music genres, which is another work we will involve in the future.

Author Contributions: Conceptualization, F.J.; methodology, F.J.; software, F.J.; validation, K.W., X.D., and W.Y.; formal analysis, F.J. and K.W.; investigation, L.Z.; resources, L.Z.; data curation, L.Z. and F.J.; writing—original draft preparation, F.J.; writing—review and editing, L.Z. and W.Y.; visualization, F.J.; supervision, L.Z.; project administration, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shaanxi Key Laboratory for Network Computing and Security Technology grant number NCST2021YB-04, the scholarship from National Natural Science Foundation of China grant number 61802301 and 211817019, Shaanxi Natural Science Foundation of China grant number 2019JQ-056.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Samples of the music are available at https://chiangfaith.github.io (accessed on 10 March 2022).

Abbreviations

The following abbreviations are used in this manuscript:

TCPS	Traditional Chinese pentatonic scale
3D-SCN	Three dimensional sequential convolutional network
BoYaTCN	Bidirectional octave your attention temporal convolutional network

References

1. Cope, D. *The Algorithmic Composer*; AR Editions, Inc.: Middleton, WI, USA, 2000 ; Volume 16.
2. Nierhaus, G. *Algorithmic Composition: Paradigms of Automated Music Generation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
3. Herremans, D.; Chuan, C.H.; Chew, E. A functional taxonomy of music generation systems. *ACM Comput. Surv.* **2017**, *50*, 1–30. [CrossRef]
4. Fernández, J.D.; Vico, F. AI methods in algorithmic composition: A comprehensive survey. *J. Artif. Intell. Res.* **2013**, *48*, 513–582. [CrossRef]
5. Briot, J.P.; Pachet, F. Deep learning for music generation: Challenges and directions. *Neural Comput. Appl.* **2020**, *32*, 981–993. [CrossRef]
6. Liu, C.H.; Ting, C.K. Computational intelligence in music composition: A survey. *IEEE Trans. Emerg. Top. Comput.* **2016**, *1*, 2–15. [CrossRef]
7. Fiebrink, R.; Caramiaux, B. The machine learning algorithm as creative musical tool. *arXiv* **2016**, arXiv:1611.00379.
8. Acampora, G.; Cadenas, J.M.; De Prisco, R.; Loia, V.; Munoz, E.; Zaccagnino, R. A hybrid computational intelligence approach for automatic music composition. In Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, 27–30 June 2011; pp. 202–209.
9. Kong, Q.; Li, B.; Chen, J.; Wang, Y. Giantmidi-piano: A large-scale midi dataset for classical piano music. *arXiv* **2020**, arXiv:2010.07061.
10. Aiolli, F. A Preliminary Study on a Recommender System for the Million Songs Dataset Challenge. In Proceedings of the ECAI Workshop on Preference Learning: Problems and Application in AI, State College, PA, USA, 25–31 July 2013; pp. 73–83.
11. Lyu, Q.; Wu, Z.; Zhu, J.; Meng, H. Modelling high-dimensional sequences with lstm-rtrbm: Application to polyphonic music generation. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.

12. Chu, H.; Urtasun, R.; Fidler, S. Song from PI: A musically plausible network for pop music generation. *arXiv* **2016**, arXiv:1611.03477.
13. Choi, K.; Fazekas, G.; Sandler, M. Text-based LSTM networks for automatic music composition. *arXiv* **2016**, arXiv:1604.05358.
14. Johnson, D.D. Generating polyphonic music using tied parallel networks. In Proceedings of the International Conference on Evolutionary and Biologically Inspired Music and Art, Amsterdam, The Netherlands, 19–21 April 2017 ; pp. 128–143.
15. Lim, H.; Rhyu, S.; Lee, K. Chord generation from symbolic melody using BLSTM networks. *arXiv* **2017**, arXiv:1712.01011.
16. Sun, F. DeepHear—Composing and Harmonizing Music with Neural Networks. 2017. Available online: https://fephsun.github.io/2015/09/01/neural-music.html (accessed on 1 September 2015) .
17. Hadjeres, G.; Nielsen, F. Interactive music generation with positional constraints using anticipation-rnns. *arXiv* **2017**, arXiv:1709.06404.
18. Waite, E.; Eck, D.; Roberts, A.; Abolafia, D. Project Magenta: Generating Long-Term Structure in Songs and Stories. Available online: https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn (accessed on 15 July 2016).
19. Mehri, S.; Kumar, K.; Gulrajani, I.; Kumar, R.; Jain, S.; Sotelo, J.; Courville, A.; Bengio, Y. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv* **2016**, arXiv:1612.07837.
20. Myburgh, J.C.; Mouton, C.; Davel, M.H. Tracking translation invariance in CNNs. In Proceedings of the Southern African Conference for Artificial Intelligence Research, Muldersdrift, South Africa, 22–26 February 2021; pp. 282–295.
21. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
22. Yang, L.C.; Chou, S.Y.; Yang, Y.H. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv* **2017**, arXiv:1703.10847.
23. Dong, H.W.; Hsiao, W.Y.; Yang, L.C.; Yang, Y.H. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
24. Bickerman, G.; Bosley, S.; Swire, P.; Keller, R.M. Learning to Create Jazz Melodies Using Deep Belief Nets. In Proceedings of the First International Conference on Computational Creativity, Lisbon, Portugal, 7–9 January 2010.
25. Lattner, S.; Grachten, M.; Widmer, G. Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints. *J. Creat. Music. Syst.* **2018**, *2*, 1–31.
26. Huang, C.Z.A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Hawthorne, C.; Dai, A.; Hoffman, M.; Eck, D. Music transformer: Generating music with long-term structure. *arXiv* **2018**, arXiv:1809.04281.
27. Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; Eck, D. A hierarchical latent vector model for learning long-term structure in music. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4364–4373.
28. Roberts, A.; Engel, J.; Raffel, C.; Simon, I.; Hawthorne, C. MusicVAE: Creating a Palette for Musical Scores with Machine Learning. 2018. Available online: https://magenta.tensorflow.org/music-vae (accessed on 15 March 2010) .
29. Chung, Y.A.; Wu, C.C.; Shen, C.H.; Lee, H.Y.; Lee, L.S. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *arXiv* **2016**, arXiv:1603.00982.
30. Boulanger-Lewandowski, N.; Bengio, Y.; Vincent, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv* **2012**, arXiv:1206.6392.
31. Hadjeres, G.; Pachet, F.; Nielsen, F. Deepbach: A steerable model for bach chorales generation. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1362–1371.
32. Bretan, M.; Weinberg, G.; Heck, L. A unit selection methodology for music generation using deep neural networks. *arXiv* **2016**, arXiv:1612.03789.
33. Luo, J.; Yang, X.; Ji, S.; Li, J. MG-VAE: Deep Chinese folk songs generation with specific regional styles. In *Proceedings of the 7th Conference on Sound and Music Technology (CSMT)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 93–106.
34. Li, J.; Luo, J.; Ding, J.; Zhao, X.; Yang, X. Regional classification of Chinese folk songs based on CRF model. *Multimed. Tools Appl.* **2019**, *78*, 11563–11584. [CrossRef]
35. Zheng, X.; Wang, L.; Li, D.; Shen, L.; Gao, Y.; Guo, W.; Wang, Y. Algorithm composition of Chinese folk music based on swarm intelligence. *Int. J. Comput. Sci. Math.* **2017**, *8*, 437–446. [CrossRef]
36. Kuo-Huang, H. Folk songs of the Han Chinese: Characteristics and classifications. *Asian Music* **1989**, *20*, 107–128. [CrossRef]
37. Liumei, Z.; Fanzhi, J.; Jiao, L.; Gang, M.; Tianshi, L. K-means clustering analysis of Chinese traditional folk music based on midi music textualization. In Proceedings of the 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 9–11 April 2021; pp. 1062–1066.
38. Zhang, L.M.; Jiang, F.Z. Visualizing Symbolic Music via Textualization: An Empirical Study on Chinese Traditional Folk Music. In Proceedings of the International Conference on Mobile Multimedia Communications , Guiyang, China, 23–25 July 2021; Springer: Cham, Switzerland; pp. 647–662.
39. Xiaofeng, C. The Law of Five Degrees and pentatonic scale. *Today's Sci. Court.* **2006**, *5* .
40. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
41. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **2017**, *42*, 1–21. [CrossRef]

42. Marcos, S.; Werner, J.S.; Burns, S.A.; Merigan, W.H.; Artal, P.; Atchison, D.A.; Hampson, K.M.; Legras, R.; Lundstrom, L.; Yoon, G.; et al. Vision science and adaptive optics, the state of the field. *Vis. Res.* **2017**, *132*, 3–33. [CrossRef] [PubMed]
43. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod Rec.* **1996**, *25*, 103–114. [CrossRef]
44. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
46. Yijun, Z. Analysis of the relationship between traditional Chinese musicology and ethnomusicology. *North. Music* **2020**, *2*, 39–40.
47. Aiqi, L. Looks at the heritage of traditional Chinese culture from the current situation of the court Yayue in China. *North. Music* **2012**, *8*, 131.
48. Xiaoqian, P. "Qing" and "Qing Yue" Theory. *Huang Zhong J. Wuhan Conserv. Music* **2012**, *1*, 75–81.
49. Zhen, Y. Rethinking of "Song lyrics originated in Yanyue"—With Mr. Li Changji. *Lit. Herit.* **2004**, *5*, 71–84.
50. Chonguang, L. *Fundamentals of Music Theory*; People's Music Press: Beijing, China, 2000; pp. 254–296.
51. Liumei, Z.; Fanzhi, J.; Jie, C.; Yi, S.; Luo, L. 3D-SCN: Three-dimensional Sequential Convolutional Networks for Music Generation of Traditional Chinese Pentatonic Scale. *J. New Music. Res.* 2022, *under review*.