Music Audio Rhythm Recognition Based on Recurrent Neural Network

Bo Chen

School of Music, Huainan Normal University, Anhui, Huainan 232038, China

Correspondence should be addressed to Bo Chen; 2018020900028@jlxy.nju.edu.cn

Received 16 January 2022; Revised 15 February 2022; Accepted 16 February 2022; Published 18 March 2022

Academic Editor: Xin Ning

Copyright © 2022 Bo Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Music rhythm detection and tracking is an important part of music understanding system and music visualization system. Based on the important position of rhythm in music expression and the wide range of multimedia applications, rhythm extraction has become an important hotspot in computer music analysis. In the field of audio recognition research, deep learning can automatically learn the features of audio and extract the rhythm of music. This paper takes music audio rhythm recognition as the main research object and carries out a series of researches with deep learning GRU neural network as the main technical support. A residual network is introduced into the GRU network model, and it is found that when the residual network is at 50 layers, the model has the highest accuracy for audio rhythm extraction. After adjusting the model parameters through experiments, this paper concludes that the average recognition accuracy of the ResNet_50-GRU model for recognizing the rhythm of the music audio in the MSD, AudioSet, and FMA data sets is 92.5%.

1. Introduction

Music is an acoustic sign, the most abstract of all human art genres, and its ideas and emotions have evolved over the centuries. The main elements of music are rhythm, melody, harmony, and sound. With the proper research and application of computer technology in the multimedia field, the multimedia process has developed rapidly. However, as the most important form of multimedia audio data, musicians perfectly combine the basic elements of music through computers, presenting a rich emotional world.

Deep learning [1] is a multilevel neural network that can recognize and simulate nonlinear mapping in specific situations and has achieved very successful results in the fields of image recognition and machine translation. Deep learning can be used as a classification tool in speech recognition research, but deeper network structures can enhance learning capabilities, and undersupervised learning, deep learning can automatically learn audio features.

Rhythm usually means the phenomenon of changes in the strength and weakness of music. Without rhythm, music loses its ability to express creativity. With the importance of rhythm in music performance and the expansion of multimedia applications, rhythm recognition has become an important focus of intelligent music analysis and has a wide range of applications in computer multimedia and other fields. At the same time, with the development of computer technology, communication, electronic technology, and multimedia, researchers will focus on the field of rhythm design using logical processes to control music production and try to establish the use of artificial intelligence for music creation.

The proposed models or algorithms for music rhythm extraction have long been a hot topic in computer analysis, and they are constantly being improved. The following are the main breakthroughs in audio rhythm recognition using a recurrent neural network: First, this paper uses a GRU neural network to build a short-term music audio rhythm extraction and recognition model. Second, in the experiment, the residual network is introduced into the GRU network model, and it is verified that the residual network can improve the model’s recognition accuracy to
some extent. Third, this paper examines the impact of various activation functions on the GRU model’s recognition accuracy.

2. Related Work

Holden proposed a real-time character control mechanism based on phase function neural network. In this network structure, the weights are calculated by a circular function using the phase as input. As the stage progresses, the system takes the previous state of the character, the geometry of the scene as input user controls and automatically generates high-quality motion to achieve the desired user controls. A new alternating update clique convolutional neural network structure (CliqueNet) can obtain deeper network, which improves the utilization of network features. To maximize the transfer of semantic information [2], Wu introduced the clique of CliqueNet. He proposed a new fully convolutional network based on the encoder-decoder structure, called CyclicNet, an alternating update network for semantic segmentation. In addition, long-hop connections and short-hop connections are added to the network to avoid vanishing gradients [2]. Granero-Molina proposed a recurrent neural network (RNN) for solving linear programming problems, which has good natural convergence and fast convergence. In order to achieve the optimal accuracy and computational complexity, an algorithm is also proposed. It presents the MATLAB-Simulink modeling and simulation verification of this recurrent neural network. The modeling and simulation results verify the theoretical analysis and effectiveness of the regression neural network for solving linear programming problems. The application of RNN to music recognition demonstrates the performance of recurrent neural networks [3]. Jin proposed an improved finite-time zero-convergence modified neural network (FTCZNN) to solve time-varying complex linear matrix formulas (TVLCM) online. To converge the error matrix to zero, the new FTCZNN uses a new design formula. The new FTCZNN converges to the TVLCM theoretical solution in finite time, according to theoretical analysis. He also developed CZN to solve the same TVLCM for comparison. The new FTCZNN has better convergence performance than the fast convergence CZNN [4]. To obtain the frequency spectrum, Ma uses a short-time Fourier transform of the music signal. The autocorrelation properties of the endpoint intensity curves were used to extract pulse code modulation (PCM) values. He proposed a multipath search and cluster analysis-based rhythm detection algorithm. That is, he proposed a multipath detection and tracking algorithm based on the clustering algorithm and incorporating the idea of multipath tracking. It eliminates the drawback that clustering algorithms necessitate the use of digital interface tools (MIDI) to assist the input in achieving the desired result. The algorithm uses a PCM signal as an input because it is more practical [5]. Based on the fusion of visual and acoustic features, Nanni et al. proposed a new and effective audio automatic classification method. They assess and compare sound’s acoustic characteristics. They then combined these characteristics into an ensemble that improved classification accuracy over more advanced methods. A different support vector machine (SVM) [6] is trained for each feature descriptor.

3. Music Audio Rhythm Recognition Based on Recurrent Neural Network

3.1. Recurrent Neural Network (RNN). RNN is a type of neural network that is very effective for data with sequence properties. It can mine time series information and semantic information in data and make use of this ability of RNN to make deep learning models make breakthroughs in solving problems in NLP fields such as speech recognition, language model, machine translation, and time series analysis [7]. The difference between RNN and fully connected neural network is that it can combine the content of the data before and after to train the model and introduces the time weight matrix, which can accurately identify the content in the context of special scenes. The RNN structure diagram is shown in Figure 1.

V and U are the parameter matrix from the hidden layer to the output layer and the parameter matrix from the input layer to the hidden layer, respectively. The training of RNN is similar to the training of traditional ANN, and the error back-propagation is used, but the parameters W, U, and V are shared. In the gradient descent algorithm, the output of each step is not only affected by the current step network but also depends on the network state of the previous steps [8].

(1) Forward propagation

RNN forward propagation is similar to the perceptron model with only one hidden layer. Assuming a sequence X of length T, the number of RNN network input layer units is A, and the number of hidden layer and output layer units is B and C. Iteratively calculate Formulas (1) and (2) from time $t = 1$ until the entire input sequence is completed. Among them, $x_{at}$ represents the value of the input unit a at time t, and $s_{at}$ and $q_{at}$ represent the value collected by the hidden unit b of the neural network at time t and the value calculated by the planning function, respectively.

$$s_{bt} = \sum_{a=1}^{A} w_{ab} x_{at} + \sum_{b'}^{B} w_{b'b} q_{b't}^{-1}, \hspace{1cm} (1)$$

$$q_{bt} = \theta_{b}(s_{bt}). \hspace{1cm} (2)$$

(2) Output layer

The output vector of the neural network is given by the output layer activation, and the input value of each output layer $c$ of the neural network is the sum of $q_{ct}$.

$$s_{ct} = \sum_{bc} w_{bc} q_{bt}. \hspace{1cm} (3)$$
The number of output layer units and the selection of the activation function should be determined according to the specific application scenario. When the number of classifications is large, the Softmax function can be used as the activation function to obtain the probability value of the classification result.

\[ P(C_c \mid x) = y_c = \frac{e^{x_c}}{\sum_{c=1}^{C} e^{x_{c'}}}. \]  

The target probability can be expressed as

\[ P(z \mid x) = \prod_{c=1}^{C} y_c^z. \]  

(3) Backward propagation

The loss function for a given RNN model is

\[ \mathcal{L}(x) = -\sum_{c=1}^{C} z_c \ln y_c. \]  

Model parameters are derived using the BPTT algorithm. BPTT is similar to standard back-propagation, BPTT contains repeated chain rules. The difference is that for a recurrent neural network, the activation function of the hidden layer from the loss function will not only affect the output layer but also affect the next moment of the hidden layer. The Formula is

\[ q_t^i = q_t^{i'}(s_b^t) \left( \sum_{c=1}^{C} q_t^{i'}w_{bc} + \sum_{b'} \sum_{b''} q_{t+1}^{i''}w_{b'b''} \right), \]  

\[ \hat{q}_t^i = \frac{\partial \mathcal{L}}{\partial s_t^i}. \]  

The complete sequence of \( q \) is iteratively calculated from time \( T \) using Formula (8), and the final derivation result of the completed parameters is

\[ \frac{\partial \mathcal{L}}{\partial w_q} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial s_t^i} \frac{\partial s_t^i}{\partial w_q} = \sum_{t=1}^{T} \hat{q}_t^i q_t^i. \]  

\[ z_t = \delta(w_t \cdot [h_{t-1}, x_t]), \]  

\[ r_t = \delta(w_r \cdot [h_{t-1}, x_t]). \]  

Then through the reset gate, calculate the candidate vector added to the hidden state of the current time step, as shown in Formula (12).

\[ \tilde{h}_t = \tanh (w \cdot [r_t \ast h_{t-1}, x_t]). \]
Finally, with the updated gate value as the weight, the candidate vector and the hidden layer state of the previous time step are mixed to obtain the output of the GRU network at time step $t$.

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t.$$  \hspace{1cm} (13)

The main reason why the GRU network can slow down the gradient disappearance phenomenon is that the gating unit acts as a "short circuit" mechanism. Through the control of the parameters of the gating unit, the previous memory is selectively retained without being erased, and the gradient is not easily attenuated when it propagates backward in the direction of the delay time axis, which greatly slows down the disappearance of the gradient.

### 3.2.2. LSTM Neural Network
LSTM is a long- and short-term memory model, which was proposed in 1997 and is also a special recurrent neural network structure. The key structure of LSTM is a unit state. This unit state has the ability to automatically add or delete information. The way to achieve these functions is to use a threshold to filter the information output by this memory unit. It includes input threshold layer and forgets threshold layer and output threshold layer. These thresholds can be self-looping parameters and can be changed according to the context. The invariance in the information flow process is guaranteed by the combination of multiple gate structures, and it is the mutual cooperation of these gate structures that the cell state can be well controlled and protected in LSTM [15–17]. The schematic diagram of its structure is shown in Figure 3.

### 3.3. Music Audio Rhythm Recognition Method

#### 3.3.1. MFCC Feature Extraction
MFCC is the most commonly used feature to solve audio recognition problems, mainly because MFCC can represent the processing results of human ears on audio. The Mel filter bank can relatively accurately describe the filtering effect of the cochlea. The Mel frequency is used in the Mel filter bank to describe the audio features, and the mapping relationship between it and the normal Hertz frequency can be described by Formula (14).

$$f = 2595 \log_{10} \left( 1 + \frac{f_{He}}{700} \right). \hspace{1cm} (14)$$

The MFCC parameter calculation method is shown in Figure 4. The main steps are preprocessing, FFT, Mel filtering, and discrete cosine transform:

Preemphasis is mainly to enhance the high frequency of the audio but also to highlight the formant. The calculation process is mainly to pass the digitally sampled signal, that is, the input signal, through a high-pass filter. The formula in the frequency domain is

$$H(z) = 1 - \frac{\alpha}{z}, \hspace{1cm} (0.9 \leq \alpha \leq 1). \hspace{1cm} (15)$$

Framing is the process of dividing the original audio into multiple short audio frames in audio signal preprocessing. The main reason is that the amplitude of the original audio changes drastically in the whole process, and the audio signal is approximately considered to be stable in a short period of time by framing. In general, the size of a frame is between 20 milliseconds and 40 milliseconds. If the frame length is set too small, it will result in too few sample points, which is not convenient for analysis. If the frame length is set too large, the signal will not be stable enough. In order to ensure the continuity of adjacent frames, adjacent frames will have a certain overlap, which also avoids the loss of critical point information.

After framing, each audio frame needs to be windowed to facilitate subsequent Fourier expansion to further analyze the audio. The advantage of windowing is that it can not only avoid the Gibbs effect but also make the global signal analysis more continuous. It can also make the original audio signal exhibit some characteristics of periodic functions and reduce the size of side lobes and spectral leakage after FFT. However, windowing will cause the loss of signal energy at both ends of the audio frame. In order to avoid missing important audio information, there is usually a partial overlap between adjacent frames. Windowing is performed on each frame of signal during calculation. There are many optional window functions, such as rectangular,
Hamming, and Gaussian windows. By using a windowing technique, the impact of performing the FFT in noninteger cycles can be minimized [18].

$$y_n(n) = z(n)y(n).$$  \hspace{1cm} (16)

Among them, $y(n)$ is the time domain signal, $z(n)$ is the window function, and $y_n(n)$ can be obtained by truncating the time domain signal by using the window function.

In order to get the frequency domain information of the signal, fast Fourier transform is performed.

$$y_i(k) = \sum_{n=1}^{N} y_i(k) \frac{1}{e^{-j2\pi kn/N}}, 1 \leq k \leq K.$$  \hspace{1cm} (17)

The modulus value of the calculated signal spectrum is squared, and the calculated power spectrum is the signal.

$$F_i(k) = \frac{|y_i(k)|^2}{N}.$$  \hspace{1cm} (18)

The Mel filter bank is composed of triangular bandpass filters. The triangular bandpass filter can have different forms, for example, the filter waveform can be of equal height, or it can change exponentially, and it can also be in the form of an inverse filter. Taking $f(r)$ to represent the center frequency of the rth triangular bandpass filter, the frequency response of the filter bank is

$$H_r(k) = \begin{cases} 
0, & k \leq f(r-1) \\
\frac{2[k-f(r-1)]}{[f(r+1)-f(r-1)] * [f(r)-f(r-1)]}f(r-1) \leq k \leq f(r) \\
\frac{2[f(r+1)-k]}{[f(r+1)-f(r-1)] * [f(r+1)-f(r)]}f(r) \leq k \leq f(r+1) \\
0, & k > f(r+1) 
\end{cases}.$$  \hspace{1cm} (19)

The discrete cosine transform actually reduces the dimensionality of the data, which is a kind of lossy compression. It is mainly aimed at some tasks that do not require high compression accuracy. The discrete cosine transform can concentrate the signal energy. $N$ is the dimension of the MFCC features to be extracted.

$$D_i = \sqrt{2 * N^{-1} \sum_{j=1}^{N} r_j \cos \left( \frac{\pi(j-0.5)}{N} \right)}.$$  \hspace{1cm} (20)

### 3.3.2. Description of Frequency Measurement Algorithm

In order to accurately measure the frequency of the output signal, the frequency measurement algorithm uses the method of measuring the cycle to indirectly convert the signal frequency, and the measurement signal cycle adopts the accumulation method to record the number of pulses $N$, the time $A$, and $B$ of the first and last pulse within 200 ms. The measurement cycle process is as follows:

First, the number of pulses is initialized, $N = 0$. Secondly, record the count value $A$ of the counter when the first pulse comes, and $N = 1$ at the same time. Then, record the count value $B$ of the counter when the next pulse comes, and $N = N + 1$, then calculate the time difference, if the result is less than 200 ms, return to the previous step, stop counting until the difference is greater than or equal to 200 ms, and then calculate the period [19, 20].

Using the evaluation algorithm, the measurement data are shown in Table 1.

| $f$ (Hz) | Theoretical value (ms) | Measured value (ms) |
|---------|-------------------------|---------------------|
| 12      | 86.36                   | 86                  |
| 16      | 74.39                   | 74                  |
| 20      | 65.56                   | 65                  |
| 24      | 52.62                   | 52                  |
| 28      | 41.84                   | 41                  |
| 32      | 36.5                    | 36                  |

The measured period is converted to obtain the corresponding signal frequency. The performance of the frequency measurement algorithm is shown in Figure 5, and the relative error is used as the performance index. The relative error calculation method is

$$\Delta T = \left| \frac{T - T_f}{T_f} \right|.$$  \hspace{1cm} (21)
As can be seen from the Figure 5, the frequency measurement algorithm takes the relative error as the performance index, and the calculated maximum relative error is 2.05%, which is about 1 ms. Therefore, for low-frequency signals, the error is negligible.

3.3.3. GRU-Based Audio Rhythm Extraction Model. The music signal will have a sudden change in energy at the general rhythm point if the music energy waveform is observed. The energy waveform of a piece of music is shown in Figure 6, and the energy difference between the rhythm and nonrhythm points can be clearly seen. As a result, the basic idea of audio rhythm extraction is that the maximum value of the sampling point frequency is compared as the rhythm point during a period of music audio. Then, by analyzing the time interval between the rhythm points, the interfering rhythm points are removed, and the maximum value detected in each second is used as the rhythm output point.

This paper uses GRU network for rhythm recognition of music audio. The specific structure is shown in Figure 7, which mainly includes audio data input, logarithmic Mel spectrogram feature extraction, feature learning and training of feature vectors by GRU network, Dropout layer, activation function for final recognition, and output rhythm recognition results. In the feature extraction step, logarithmic Mel spectrogram and CQT spectrogram are selected as system features. The logarithmic Mel spectrogram is based on MFCC extraction, and the step after DCT is removed in the MFCC extraction step. Then, two steps are added, namely, obtaining the energy spectrum and performing logarithmic operations on the energy spectrum. The choice of activation function will be compared in the experimental part to compare the recognition accuracy of models using different activation functions.

4. Music Audio Rhythm Recognition Experiment Based on GUR Recurrent Neural Network

4.1. Experimental Data and Settings. In the experiment of music audio rhythm recognition method based on GUR recurrent neural network, this paper chooses to use MATLAB for simulation. This paper selects three public music audio data sets, namely MSD (Million Song Dataset), FMA (Free Music Archive), and AudioSet. Taking MSD as an example to introduce the music audio data set, MSD is similar to a resource integration platform, which collects data from 7 well-known and authoritative music communities such as SecondHandSongs dataset and Last.fm dataset. In addition to the original data of major music websites, MSD also conducted necessary analysis and extraction on them. AudioSet is an extended ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips extracted from YouTube videos. FMA is a data set for music analysis, 1000 GB in size.

Before conducting the experiment, each data set should be preprocessed, a certain number of sample data should be selected, divided into training set and test set according to the ratio of 3: 1, and the music audio duration of the selected samples should be unified. It also performs Mel filtering to obtain a Mel spectrogram of the audio in the data
The results of preprocessing on three music audio data sets are shown in Table 2.

After the frame length and frame shift of the preprocessing are fixed, the size of the audio duration determines the value of the first dimension in the two-dimensional matrix, and the Mel filter fixes the value of the second dimension. Taking the MSD data set as an example, the size of the Mel spectrogram is 498 $\times$ 64, and 498 represents the number of audio frames divided by 8 seconds of audio with a frame length of 25 milliseconds and a frame shift of 10 milliseconds. 64 represents the energy information obtained by the Mel filter for each frame after FFT and other operations.

### 4.2. The Experimental Results of Audio Rhythm Recognition by Recurrent Neural Network

The first step in the experiment is to look at the recognition accuracy of traditional RNN neural networks, GRU neural networks, and LSTM neural networks, as well as the effect of the Lag Window size, or the size of the sliding window used for training and prediction, on the error value. The relative error of the model's recognition of audio rhythm points determines the accuracy value. The FMA sample data set is the data set used in this step. Figure 8 depicts the experimental results. Based on the experimental results, it can be concluded that the Lag Window value corresponding to each RNN’s minimum error is not consistent, and that determining a general Lag Window parameter that is suitable for most RNN models is difficult. And the neural network rhythm recognition model's relative error value SMAPE has a trend of decreasing and then increasing, indicating that the size of the Lag Window and the recognition accuracy of the cyclic neural network music audio rhythm recognition model are not completely correlated. The main reason is that if the selected Lag Window value is larger, the range of identification data of each node is enlarged. That is to say, a part of data that is far away from the time to be recognized will be introduced, which will bring redundant information or noise, which will lead to loss of recognition accuracy. In addition, the increase in the dimension of model input data will also reduce the recognition efficiency of the model. Then, if the selected Lag Window value is small, the recognition model does not recognize enough information, and it is impossible to learn the peak change pattern within the sequence, and it is difficult to make a relatively accurate recognition. This also shows that the Lag Window, whether large or small, will have an adverse effect on the prediction results, that is, each prediction model is very sensitive to this parameter. Then, for the three cyclic neural networks used in the comparative experiments in this paper, within a certain range, the error mean of the GRU and LSTM models is smaller than that of the RNN model. Compared with GRU and LSTM, the mean error of GRU will be relatively small.

The second step of the experiment is to introduce a residual network (Res Net) in the GRU music audio rhythm recognition model. The idea of ResNet is to assume that we involve a network layer and there is an optimized network layer, so often the deep network we design has many network layers as redundant layers. Then, we hope that these redundant layers can complete the identity mapping to ensure that the input and output through the identity layer are exactly the same. The specific layers are identity layers, which will be judged by yourself during network training. As the depth of the network increases, the model may degenerate. That is to say, if there are redundant network layers in the neural network model, the recognition accuracy may be lower than that of the model with fewer layers, so it

| Type    | Number of samples | Training set | Test set | Audio duration(s) | Mel map size |
|---------|------------------|--------------|----------|-------------------|--------------|
| MSD     | 1500             | 1125         | 375      | 8                 | 498 $\times$ 64 |
| AudioSet| 480              | 360          | 120      | 6                 | 398 $\times$ 64 |
| FMA     | 600              | 450          | 150      | 6                 | 398 $\times$ 64 |
Figure 8: Relative error changes under different Lag Window.

Figure 9: The effect of different layers of residual network on the recognition accuracy of GRU model.

Figure 10: Recognition accuracy of different activation functions.
is necessary for the model to train the redundant layers to be identical in the process of training the network model. Mapping layers, that is, data passing through these layers does not change the input and output values. In this step of the experiment, the rhythm recognition accuracy experiments were carried out on the MSD, AudioSet, and FMA data sets, respectively. The layers of the residual network were 20, 50, 80, and 110 layers, respectively. The experimental results are shown in Figure 9. It can be seen that in the experiments on the three types of audio data sets, the rhythm recognition accuracy is the highest when the residual network is 50 layers, and when the residual network layer is more than 50 layers, the increase in the number of residual network layers reduces the recognition accuracy. This may be because as the number of network layers deepens, a certain overfitting problem occurs, so the training accuracy cannot be improved, and the training time is also increasing. In summary, this paper chooses to introduce a residual network structure with 50 layers.

The experiment next studies the relationship between the choice of activation function in the ResNet_50-GRU model and the model’s rhythm recognition accuracy. The activation function is to introduce nonlinear factors into neurons, so that the neural network can approximate any nonlinear function arbitrarily, so that the neural network can be applied to many nonlinear models. The activation functions selected for comparative experiments are Softmax function, RELU function, and Tanh function, and experiments are also performed on different audio data sets. The experimental index is the recognition accuracy of audio rhythm. The training set experiment is carried out first, and then, the test set experiment is carried out. The experimental results are shown in Figure 10. It can be seen from the experimental results that the recognition accuracy of the ResNet_50-GRU recurrent neural network recognition model of the Softmax activation function is higher than that of the RELU activation function and the Tanh activation function model. This may be because the distribution of audio tempo over time may be closer to a discrete probability distribution, and the Softmax function is essentially a discrete probability distribution corresponding to multiclassification tasks. Taking the output recognition results of the test set as an example, compared with the model without the activation function, the recognition accuracy of the ResNet_50-GRU recurrent neural network recognition model with Softmax activation function on the MSD, AudioSet, and FMA audio data sets has been improved by 5.4%, 3.3% and 7.2%. The recognition accuracy on the FMA under the three data sets is relatively low, which may be due to the fact that the peak time of the audio signal of the experimental sample selected in this data set is relatively close, and the discrimination between the rhythm points of the audio is relatively low.

When evaluating the recognition system, in addition to statistical recognition accuracy, confusion matrix can also be used for intuitive description. Confusion matrix, also known as error matrix, is a visualization tool that can reflect the accuracy of recognition from different perspectives. For the ResNet_50-GRU model whose activation function is the Softmax function, the rhythm extraction and recognition tests are performed on the fast three music pieces. The energy envelope, frequency sampling points, and actual rhythm points of its audio is shown in Figure 11. The rhythm points in this segment of audio are represented by A-D, assuming that the identified rhythm points are represented by 1, and the nonrhythm points are represented by 0. The specific confusion matrix results are shown in Table 3.

The rhythm points in this segment of audio are represented by A-D, assuming that the identified rhythm points are represented by 1, and the nonrhythm points are represented by 0. The specific confusion matrix results are shown in Table 3.

The table shows that the majority of the rhythm points in the audio can be accurately identified, but the model does not recognize the timing node as a rhythm point for the fourth rhythm point in this audio segment. The reason for this analysis could be that the rhythm point’s peak value is

| Actual distinguish | A   | B   | C   | D   |
|-------------------|-----|-----|-----|-----|
| A                 | 11  | 0   | 0   | 0   |
| B                 | 0   | 11  | 0   | 0   |
| C                 | 0   | 0   | 11  | 0   |
| D                 | 0   | 0   | 0   | 10  |
not high enough to show the degree of discrimination, or the sampling frequency is too low, and the model incorrectly identifies the rhythm point as a nonrhythm point. Finally, the model parameters are adjusted appropriately, and the ResNet 50-GRU model is used to train and test all sample data sets; the final rhythm point recognition accuracy of all audio data is 92.5 percent, indicating that the model has a good recognition effect.

5. Discussion

The main research direction of this paper is to study the rhythm recognition of music audio based on cyclic neural network. The main technical and theoretical supports are cyclic neural network algorithm and audio rhythm extraction theory. The audio rhythm recognition method mainly includes the preprocessing and feature extraction methods of audio data, and the recurrent neural network is used to learn the extraction rules of audio rhythm and then perform adaptive analysis on the audio. This paper firstly reviews the related technical principles. In the introduction of cyclic neural network, it mainly includes the basic principles and related content introduction of traditional RNN neural network. This paper also introduces the variant models of traditional RNN, GRU neural network, and LSTM neural network. Both GRU and LSTM are sequence processing models based on gate control units. In the introduction of related principles of music audio rhythm recognition method, the main content is elaborated and analyzed around MFCC feature extraction and frequency measurement algorithm, and based on the previous theory, an audio rhythm extraction model based on GRU neural network is proposed.

The article’s subsequent experimental section is divided into two subsections. The first subsection provides an overview of the experimental data and basic settings, as well as information on the public music audio data sets used in this study and the preprocessing process. The analysis of the specific experimental results, which is divided into four stages, is the second subsection. The GRU model chosen in this paper is compared to traditional RNN and LSTM models in the first stage. The results show that GRU has a higher rhythm recognition accuracy than the other two, and that the influence of the size of the Lag Window on the model recognition accuracy is not purely linear. The GRU model’s influence on the number of residual network layers is investigated in the second stage. After conducting tests, it was discovered that the model’s recognition accuracy is highest when the number of residual network layers is 50 layers. The third stage of experiments is to study the relationship between different activation functions and GRU audio rhythm recognition accuracy. The experimental results show that under the Softmax function, the model’s ability to analyze and identify each data set is better than the other two activation functions. The fourth stage experiment is a specific verification experiment. It identifies an audio frequency of a fast three music rhythm and gives the signal envelope and experimental matrix of the audio frequency. After analyzing the results in this paper, the model parameters are adjusted, and finally the recognition accuracy of the ResNet_50-GRU model for all sample data sets is calculated.

6. Conclusions

Recurrent neural networks are very effective for meaningful processing of data with sequential characteristics and have made major breakthroughs in problems in the field of NLP such as language modeling and speech recognition. This paper is a research on the recognition of music audio rhythm based on recurrent neural network. This paper summarizes the RNN and its variant neural network model, conducts experimental analysis on the audio data set with a certain sample size, and has made certain research progress, but there are still many shortcomings. For example, in the experiment of the influence of Lag Window size on model recognition accuracy, due to the limitation of sample size, an optimal Lag Window size value has not been researched. And the number of residual network layers selected later may only show the best effect in the model of this paper, because there are many influencing factors, and it does not have good universality.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author does not have any possible conflicts of interest.

References

[1] X. B. Jin, W. T. Gong, J. L. Kong, Y. T. Bai, and T. L. Su, “A Variational Bayesian and deep network with data self-screening layer for massive time-series data forecasting,” *Entropy*, vol. 24, no. 3, p. 335, 2022.

[2] Y. Ding, X. Zhao, Z. Zhang, W. Cai, N. Yang, and Y. Zhan, “Semi-supervised locality preserving dense graph neural network with ARMA filters and context-aware learning for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2021.

[3] J. M. Hernández-Padilla, J. Granero-Molina, V. V. Márquez-Hernández, F. Suthers, O. M. López-Entrambasaguas, and C. Fernández-Sola, “Design and validation of a three-instrument toolkit for the assessment of competence in electrocardiogram rhythm recognition,” *European Journal of Cardiovascular Nursing*, vol. 16, no. 5, pp. 425–434, 2017.

[4] J. Jin, “An improved finite time convergence recurrent neural network with application to time-varying linear complex matrix equation solution,” *Neural Processing Letters*, vol. 53, no. 1, pp. 777–786, 2021.

[5] S. Ma, “Music rhythm detection algorithm based on multipath search and cluster analysis,” *Complexity*, vol. 2021, no. 4, 2021.

[6] L. Nanni, Y. M. Costa, D. R. Lucio, C. N. Silla Jr., and S. Brahnam, “Combining visual and acoustic features for audio classification tasks,” *Pattern Recognition Letters*, vol. 88, pp. 49–56, 2017.
[7] X. B. Jin, W. T. Gong, J. L. Kong, Y. T. Bai, and T. L. Su, "PFVAE: a planar flow-based variational auto-encoder prediction model for time series data," Mathematics, vol. 10, no. 4, p. 610, 2022.

[8] W. Tang, J. Y. Gao, X. Y. Ma, C. H. Zhang, L. T. Ma, and Y. S. Wang, "Application of recurrent neural network in prognosis of peritoneal dialysis," Beijing Da Xue Xue Bao. Yi Xue Ban = Journal of Peking University. Health Sciences, vol. 51, no. 3, pp. 602–608, 2019.

[9] G. Bao, Y. Y. Chen, S. Y. Wen, and Z. C. Lai, "Stability analysis for memristive recurrent neural network and its application to associative memory," Acta Automatica Sinica, vol. 43, no. 12, pp. 2244–2252, 2017.

[10] E.-A. H. M. Le JA, M. Allali, and D. C. Struppa, "Application of recurrent neural networks for drought projections in California," Atmospheric Research, vol. 188, pp. 100–106, 2017.

[11] X. Wang, X. Chen, M. Yan, and M. Chang, "Constitutive model for ratcheting behavior of Z2CND18.12N austenitic stainless steel under non-symmetric cyclic stress based on BP neural network," Steel and Composite Structures, vol. 28, no. 5, pp. 517–525, 2018.

[12] A. Sharma, S. K. Raijput, and S. K. Soni, "Cyclic high temperature oxidation behaviour of bare and NiCr coated mild steel and low alloyed steel," Materials Today: Proceedings, vol. 5, no. 9, pp. 18433–18441, 2018.

[13] P. Lin, R. Ratnam, H. Sankari, and A. Garg, "Mechanism of microstructural variation under cyclic shearing of Shantou marine clay: experimental investigation and model development," Geotechnical and Geological Engineering, vol. 37, no. 5, pp. 4163–4210, 2019.

[14] M. Zhang, D. Wu, and R. Xue, "Hourly prediction of PM2.5 concentration in Beijing based on bi-LSTM neural network," Multimedia Tools and Applications, vol. 80, no. 16, pp. 24455–24468, 2021.

[15] B. Sri, S. Prasetyono, and L. Kusuma, "Estimation of Lead acid battery capacity using pulse voltammetry cyclic and neural network method," International Journal of Computer Applications, vol. 172, no. 6, pp. 5–8, 2017.

[16] Y. Garip and Z. Garip, "Application of artificial neural network for prediction of the cyclic oxidation behavior of electrical resistance sintered gamma-TiAl intermetallics," Archives of Metallurgy and Materials, vol. 66, no. 2, pp. 581–591, 2021.

[17] S. H. Park, D. Yoon, S. Kim, and Z. W. Geem, "Deep neural network applied to joint shear strength for exterior RC beam-column joints affected by cyclic loadings," Structure, vol. 33, no. 1, pp. 1819–1832, 2021.

[18] W. Liu, Q. Wang, and Q. Guo, "Automatic radar waveform recognition based on neural network," Control and Intelligent Systems, vol. 46, no. 2, pp. 92–96, 2018.

[19] A. N. Omel'Chenko and N. E. Makarchuk, "fMRI visualization of functional patterns of neural networks during the performance of cyclic finger movements: age-related peculiarities," Neurophysiology, vol. 49, no. 2, pp. 1–12, 2018.

[20] A. M. Thu, M. Song, S. Wu, A. Sheng, X. Chen, and X. Wang, "Artificial neural network prediction and mechanism analysis for migration of environmental contaminant cyclic organosiloxane oligomer from silicone rubber," Industrial & Engineering Chemistry Research, vol. 58, no. 25, pp. 11093–11100, 2019.