Design of intelligent evaluation algorithm for music word and song matching based on neural network

Qilin Song, Chunqiu Wang*
College of Music and Dance, Huaihua University, Huaihua, China

ABSTRACT

By collecting the listener's response to the generated music word and song matching degree, such as heartbeat, pulse and skin conductance, such as evaluation score, preference or physiological signal; Then use the collected information to evaluate the generated music works. How to evaluate the performance of many algorithms is a difficult problem, not only because of the complexity of music data itself, but also because the query requirements of users in different application environments are very different. For a piece of music, rhythm is very important. In this regard, this paper attempts to design an intelligent evaluation algorithm for music word and song matching based on NN (Neural Network), and generate TopN recommendations for target users by calculating the similarity between user preference features and music potential features. At the end of the simulation experiment, it is further proved that the algorithm proposed in this paper is convergent in the iterative process, and the accuracy in the training set has reached 93.4%. When the lyrics are looked at alone, the lyrics reflect the sad emotion. But the lyrics combined with the melody can be found that the rhythm of the song is very strong. Through the experiment, the accuracy in the test set has reached 86.2%, which shows that the accuracy of the algorithm proposed in this paper is considerable.

Keywords: Neural network; Matching degree of music lyrics; Intelligent evaluation algorithm

1. INTRODUCTION

The evaluation of the matching degree of music lyrics is an objective evaluation for the matching degree of music lyrics. It can be found that, in fact, after the exchange of emotional music words and songs, counterexample music is not matched because of emotional inconsistency on the one hand. But more specifically, the overall rhythm of lyrics and melody is inconsistent. For the application of music lyrics matching, the user wants to input a humming that may contain some errors, and the system returns the song he wants to query in the shortest possible time, instead of returning many songs that are similar to his humming input to some extent [1-2].

How to evaluate the performance of many algorithms is a difficult problem, not only because of the complexity of music data itself, but also because the query requirements of users in different application environments are very different. For a piece of music, the rhythm is very important [3]. Among them, lyrics and melodies are often in rhythm, such as the tone of pronunciation; Or for short and fast melody fragments, it often corresponds to lyrics that can be pronounced shorter, rather than complicated lyrics. Therefore, this paper tries to give an accurate and objective intelligent evaluation by comprehensively considering the emotion of lyrics and the rhythm relationship between lyrics, expounds the intelligent evaluation algorithm of music lyrics matching degree based on NN, collects the historical behavior information of music users, and constructs the user preference model by using the method of matrix decomposition of hidden semantic model; Then, the audio resources in the system are preprocessed, and the Mel spectrogram which can represent the music characteristics is extracted. Then, the NN is trained to obtain the regression model for the prediction of the potential characteristics of music, and the users and music are projected into a shared hidden space[4-5]. Finally, TopN recommendation is generated for the target user by calculating the similarity between the user's preference characteristics and the potential characteristics of music. Only by designing a reasonable musical expression form can we learn the matching degree and rhythm characteristics of lyrics and songs.

*Email: 554342852@qq.com

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik, Tao Shen, Proc. of SPIE Vol. 12779, 127791P · © 2023 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689464

2. MUSIC DATA REPRESENTATION

2.1 Construction of counterexample of music emotion

Music emotion is an art form organized according to time. Music is composed of melody, harmony, rhythm and other elements. Depending on the style or type, some of these elements may be emphasized or ignored. The range of music performance is wide, including various vocal music skills from singing to rap, and the performance skills of different kinds of instruments. Users can define labels and music understanding through their own understanding of music. Due to the inseparable relationship between music and emotion, a variety of labels diagnose the existence of a large number of emotional labels in the user's music labels from different angles. These labels either express the feeling of joy and joy, or express the feeling of sadness and loss [6]. As shown in Figure 1, the music emotion model is established, and music is divided into five categories from the perspective of emotion. Adjacent categories have certain emotional similarity, and emotions can be converted on adjacent categories. Relative to the two emotional categories, the emotional connotation is opposite.

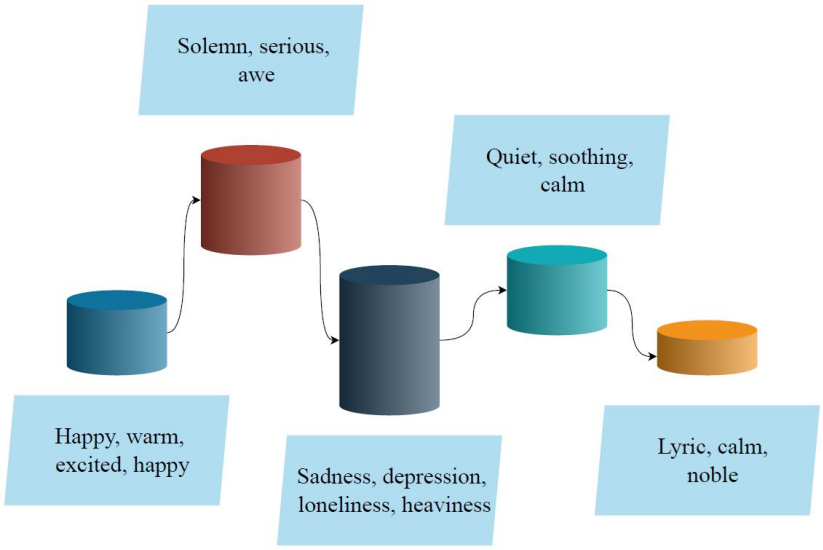

Figure 1. Music emotion model

In the process of emotional calibration, every kind of emotional music should have distinct emotional expression as much as possible. In traditional emotional music, there is not no noise. Many percussion instruments without tones, such as military drums or sand hammers, use noise to produce tones without any perceptible tones. Non-tone percussion instruments are usually used to provide rhythm or embellishment, and their sound has nothing to do with the melody and harmony of music. That is, when listening to happy music, it is easy to distinguish the music as happy music without ambiguity. At present, emotion-based music retrieval mainly obtains the emotional information of music from its own style, melody, rhythm, timbre, etc., without using a large number of emotional tag information in user tags [7].

2.2 Construction of counterexample of music rhythm

Under the same melody segment, the corresponding lyrics are similar enough, that is, they are consistent with the lyrics in rhythm. In this paper, the original rhythm connection is broken through the disorder of lyrics, so as to realize the construction of rhythm counterexample. In order to eliminate the user's humming errors, the general retrieval algorithm uses the relative features of the music, that is, the pitch difference and the sound length ratio of the music notes as the music feature value sequence. Usually, each user has his own preference for music, for example, Xiao Zhang likes quiet

and soothing songs. If a music happens to have the elements that the user likes, then he can recommend it to him [8]. Each user has different preferences for different elements, and each piece of music contains different elements.

In the western musical system, there are 12 steps in each octave in ascending order at intervals of semitones, and the interval distance of each semitone can be divided into 100 steps more accurately. For the disorder with the minimum granularity of lexical chunks, it will destroy the basic word structure, such as singing lexical chunks continuously. Such a lyric sequence is unreasonable, so it should not be out of order in this way. For humming query, it is very difficult to actually collect a number of user queries that can support performance evaluation, so it is necessary to construct a considerable number of query inputs. The construction method is that the program randomly selects a piece of music in the database, intercepts a melody from the music, and inserts certain errors for simulation [9].

2.3 Music lyrics and melody expressions

Because this paper uses English music data set, while English is phonological, in music, the suffix pronunciation of words has an impact on the rhythm of music itself. Therefore, for lyrics in music, in order to better capture the meaning and pronunciation of lyrics. It is expressed in a distributed form, and the words are divided into smaller chunks. The word coding and chunk coding are considered to fully extract the pronunciation features of English lyrics. The oboe, as the tuning standard, also produces pitch deviation, which results in the fact that the pitch frequency played by the real orchestra is generally different from that of the standard instrument, and the offset of all instruments is the same [10].

In the fields of music information retrieval and music recognition, music lyrics and melodies are very important. Similarly, for the music recommendation system in this paper, extracting effective music melody features is also a key step. Generally, musical melody features can be roughly divided into two categories: time-domain features and frequency-domain features. A good musical melody feature representation method not only needs to be able to effectively represent the lyrics and melody, but also needs to be simple and efficient. For music melody, the note is converted into pitch, length and rest time to represent, and the result of the note coding vector corresponding to the AA word block is shown in Table 1.

Table 1. Melody coding results

	A	**mer**	**wom**	**on**	**stay**
Musical note	B4	A5	A5	G5	B4
Rest	0	0	0	0	1
Sound length	0.4	0.5	0.5	0.4	0.76

In order to express the music melody more accurately, and at the same time, it can accommodate the common errors of music melody in pitch difference and tone length ratio. We can classify the pitch difference or tone length ratio more finely. By obtaining the statistical characteristics of the pitch offset of all instruments in the target music in the spectrum, we can estimate the pitch offset of the whole music and correct the spectrum accordingly. Here we say that the accuracy of pitch offset is 10 points.

3. RESEARCH ON INTELLIGENT EVALUATION ALGORITHM OF MUSIC WORD AND SONG MATCHING BASED ON NEURAL NETWORK

3.1 Algorithm design

In the context of lyrics without melody, it is easy to ignore the sense of rhythm in the lyrics. In music, the sense of rhythm is very important for the expression of emotion. As this article puts forward, emotion is largely expressed through music rhythm. Taking emotion and rhythm into consideration, the algorithm model of word and song matching is designed as shown in Figure 2.

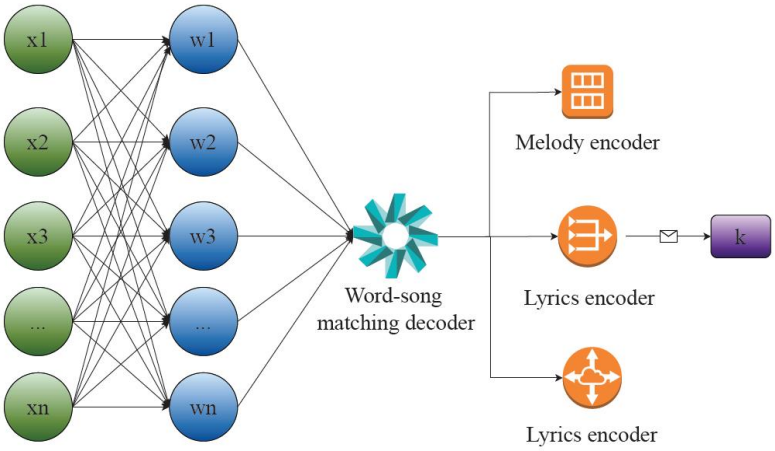

Figure 2. Word and song matching algorithm

The input layer neurons and hidden layer neurons of the word and song matching algorithm are the weights of the network. Each neuron has input and generates a single output to one or more other neurons. On average, one note is missing per paragraph. When an update error occurs, one or more errors may occur per paragraph. By learning a deep nonlinear network structure and representing a large amount of data related to users and items, we can effectively obtain the deep feature representation of users and items. Most of them are errors in the same direction as the melody outline. Based on this, recommendation by integrating traditional recommendation methods can alleviate the problems existing in the traditional system, such as data sparsity and cold start, to a certain extent.

According to the potential characteristics of music predicted by the intelligent evaluation algorithm and combined with the user preference characteristics, the matching degree between users and music is calculated, and finally the recommended list of music objects that users may be interested in is generated. The realization of the whole system function mainly includes two main processes: regression model training and prediction recommendation. The layer receiving external data is the input layer. The layer that produces the final result is the output layer. There is a hidden layer between the input layer and the output layer. There may be one or more hidden layers, or no hidden layers at all. In the memory module, the actual distribution of attention in the input can be reflected by the attention weight. The calculation process of attention weight is expressed as follows:

$$z_i^l = f_i m, f_i q, f_i - m \tag{1}$$

$$g_i^l = \frac{\exp\left(Z_i^l\right)}{\sum_{k=1}^{N} \exp} \tag{2}$$

In equations (1)-(2), the superscript l indicates the l layer, the subscript i indicates the i fact, and z_i^l and g_i^l respectively indicate the interaction characteristics between the first fact and question on the l layer and the memory information on the i layer, the attention weight corresponding to the i fact and the probabilistic attention weight.

Assume that the input signal is x_i, and its internal structure is defined as:

$$u_i = \sigma\left(W x_i + U^u h_{i-1}\right) \tag{3}$$

$$h_i = u_i h_i + \left(1 - u_i\right) h_{i-1} \tag{4}$$

In equations (3)-(4), a selection process of the current hidden state and the historical hidden state is controlled by the attention weight g_i.

With the in-depth study of the algorithm, people have improved this algorithm and put forward many effective optimization algorithms one after another, among which gradient descent method and alternating non-negative least square method are commonly used, so I won't go into details here. Generally speaking, NN provides a new way for large-scale data decomposition, and compared with the traditional matrix decomposition method, it is simpler and more efficient, has better interpretability and requires less storage space. When calculating the context vector of each layer from the input story, we can not only consider the attention weight distribution of the input facts at each moment, but also consider the position information of the input facts. In order to better test the fault-tolerant ability of the algorithm, a segment with a length of 15 notes is randomly intercepted from the database, and 1~3 errors opposite to the melody contour are randomly added to each intercepted melody segment, where the position and error type of each one are also random, as the input of each algorithm, to test the performance of the algorithm.

3.2 Discussion of experimental results

In this chapter, the performance test experiments for the calculation of editing distance, DTW and OSCM methods are performed under five conditions with data volume of 4500, 9000, 18000, 36000 and 72000 songs respectively The algorithm model uses zero value to initialize the hidden state of melody encoder and lyrics encoder. The training process uses the Adam optimizer with a learning rate of 0.01 to calculate and update the back-propagation gradient, and completes the iterative training through the loss function. The change of iteration process loss is shown in Figure 3.

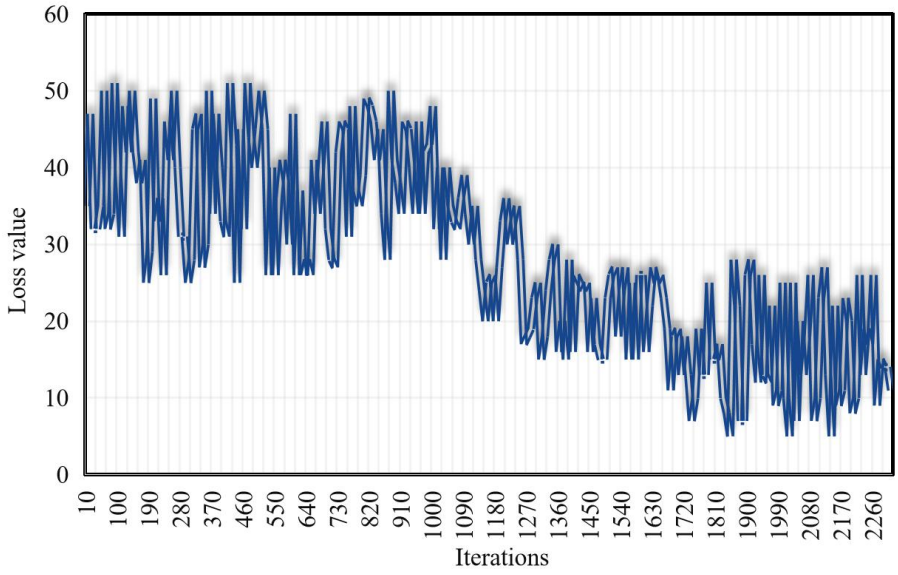

Figure 3. Loss change.

In the case of different data volumes, it is almost 98%. When users hum only two errors in the opposite direction of the melody contour, the hit rate of the top 10 is also about 90%. Therefore, it is a query algorithm suitable for large humming retrieval systems.

Considering that when designing the evaluation data set, the number of notes of each song is only intercepted to 100 notes, the average time required for each query of the dynamic planning algorithm and dynamic time warping algorithm for editing distance should be more than 3 times of the experimental time. In Figure 3, the loss value of each batch is the sum of the loss of batch samples, and the number of sample fragments of each batch is 64; The accuracy change of the iteration process is shown in Figure 4.

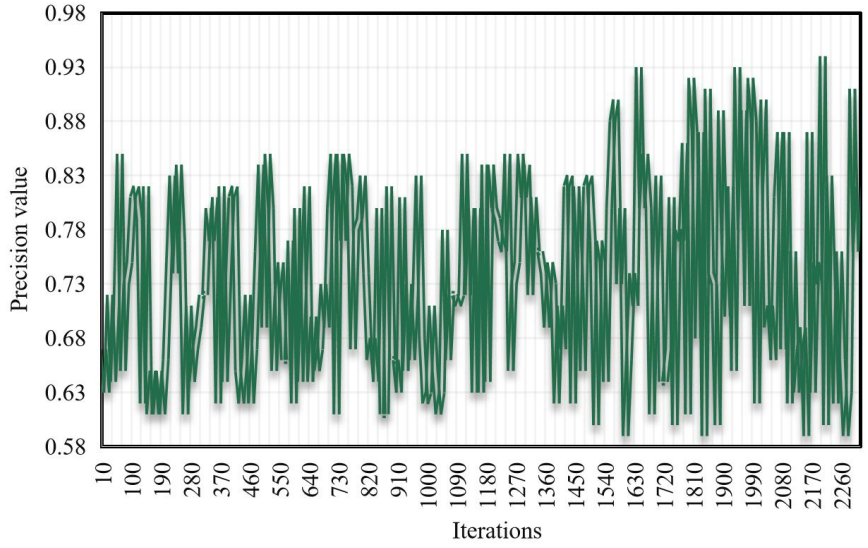

Figure 4. Accuracy change

Through the above experiments, we can find that the melody of music lyrics reflects more cheerful emotions. In the iterative process, it converges, and the accuracy in the training set reaches 93.4%. When looking at the lyrics alone, the lyrics reflect sad feelings, but the lyrics combined with the melody can be found that the song has a strong sense of rhythm, and the accuracy in the test set reaches 86.2%. Therefore, the overall rhythm is consistent, and it is a music with matching lyrics and songs. If we don't consider emotion and rhythm at the same time, and simply pass through emotional dimension, we will think that the lyrics and songs of the music don't match.

4. CONCLUSIONS

In reality, we can't ignore the impact of various physical and environmental conditions on the musical instrument. The musical instrument cannot always maintain its ideal state in the process of preservation, and the various pitches it produces may not exactly correspond to the standard frequency. The categories of songs are also limited to pop songs. In the future, the categories of people participating in the humming input test and humming songs will be expanded, and a more representative user error model will be established, so that the comparison of algorithms has a more solid foundation. The simulation experiment proves that the algorithm proposed in this paper converges in the iterative process, and the accuracy in the training set reaches 93.4%. When looking at the lyrics alone, the lyrics reflect the sad emotion. However, the lyrics combined with the melody can be found that the rhythm of the song is very strong. Through the experiment, the accuracy in the test set reaches 86.2%. It can be seen that the accuracy of the algorithm proposed in this paper is considerable. If the NN based music word and song matching degree includes the candidate selection of possible fundamental frequency and harmonic components, most of these candidates will have the same pitch offset. The normal distribution of music words and songs is introduced to estimate the offset of all candidates, and the parameters of the normal distribution represent the maximum possibility of the offset distribution.

REFERENCES

[1] Yan, F., Music Recognition Algorithm based on T-S Cognitive Neural Network. Translational Neuroscience, vol. 10, no. 5, pp. 11-19 (2021).
[2] Wu, R., Research on automatic recognition algorithm of piano music based on convolution neural network. Journal of Physics: Conference Series, vol. 1941, no. 1, pp. 012086-012099 (2021).
[3] Zhang, K., Music Style Classification Algorithm Based on Music Feature Extraction and Deep Neural Network. Wireless Communications and Mobile Computing, vol. 20, no. 4, pp. 1-7 (2021).

[4] Zheng, Y., The use of deep learning algorithm and digital media art in all-media intelligent electronic music system. PLOS ONE, vol. 15, no. 10, pp. 28-64 (2020).
[5] Deng, Y., Xu, Z., Zhou, L., et al., Research on AI Composition Recognition Based on Music Rules. vol. 58, no. 13, pp. 37-55 (2020).
[6] Baxter, M., Ha, L., Perfiliev, K., et al., Context-Based Music Recommendation Algorithm Evaluation. vol. 38, no. 11, pp. 19-33 (2021).
[7] Huang, C., Shen, D., Research on Music Emotion Intelligent Recognition and Classification Algorithm in Music Performance System. Hindawi Limited, vol. 48, no. 19, pp. 27-56 (2021).
[8] Sun, S., A College Music Teaching System Designed Based on Android Platform. Hindawi Limited, vol. 66, no. 21, pp. 29-53 (2021).
[9] Wang, D., Guo, X., Research on Intelligent Recognition and Classification Algorithm of Music Emotion in Complex System of Music Performance. Complexity, vol. 20, no. 22, pp. 36-58 (2021).
[10] Zhang, L., Tian, Z., Research on the recommendation of aerobics music adaptation based on computer aided design software. Journal of Intelligent and Fuzzy Systems, vol. 18, no. 2, pp. 1-12 (2021).