Note Extraction and Recognition Analysis Based on Music Melody Features

Yanmin Liu

[1]School of Music, Suzhou University, Suzhou, Anhui 234000, China

E-mail: liuym1990@outlook.com

Keywords: music, melody, note, convolutional neural network

Received: May 13, 2024

This paper extracted and recognized the notes in the music based on the features of the music melody. Firstly, the melodic features and Mel-frequency cepstral (MFCC) features were extracted from the music signal and then combined. A convolutional neural network (CNN) was used as a classifier for note classification and recognition in the music signal. The CNN adopted a structure with three convolutional layers and three pooling layers. The melody features and MFCC features were used to extract convolutional features through convolution kernels in the convolutional layers, followed by compressing these features in the pooling layers. Finally, the note recognition results were outputted in the output layer. Then, simulation experiments were performed using a self-built music library. The performance of the algorithm was tested under different MFCC feature dimensions and CNN activation functions. The algorithm was also compared with the dynamic time warping (DTW) algorithm and the CNN algorithm without music melody features. The results showed that the proposed algorithm had the best performance when the MFCC feature dimension was set to 24, and the CNN activation function was sigmoid; under such conditions, the F-measure was 96.7%. The performance of the proposed algorithm was the best, regardless of whether it recognized the single-note or multi-note music. The precision for recognizing single notes was 98.3%, the recall rate was 96.5%, and the F-measure was 97.4%. For recognizing multiple notes, the corresponding values were 93.0%, 92.5%, and 92.7%, respectively. However, the performance of the three algorithms was reduced when recognizing multi-note music.

Povzetek: Predstavljen je algoritem za prepoznavo glasbenih not z uporabo konvolucijskih nevronskih mrež in kombinacijo melodičnih ter MFCC lastnosti.

1 Introduction

Music is an important part of human culture, and melody is one of the core elements of music [1]. In digital music processing, effectively extracting and recognizing music melody features and converting them into note text is of great significance for understanding music content, music recommendation, music copyright protection, and other aspects [2]. In terms of music creation, extracting and recognizing the notes in the music melody can also assist the author in creating music better [3]. The melody characteristics of music include pitch, rhythm, intensity, etc., and the above characteristics can be used to identify the notes in the music melody better. The related researches are shown in Table 1. In these studies, different algorithms were used to recognize the notes in music signals. Some of them employed neural networks for note recognition, while others utilized optimization algorithms to improve the efficiency of recognition algorithms. This paper starts with the features of music signals and combines the melody features with conventional MFCC features to form new music signal features. Then, a CNN algorithm is used to recognize the notes based on these new music signal features. Using complex convolution kernels in the CNN algorithm can consider both local and global features, thereby enhancing the accuracy of note recognition.

Table 1: Related studies.

Author	Research content	Result
Tamboli et al. [4]	They designed a note detection method based on a classification framework by using an optimization-based neural network	Simulation experiments verified the effectiveness of this approach.
Cinar et al. [5]	They proposed a new time series segmentation	The simulation experiment results showed

	framework based on a hierarchical linear dynamical system and tested its performance in detecting monophonic and polyphonic notes.	that it was more accurate than the current state-of-the-art methods.
Nazar et al. [6]	They used the parallel bat algorithm to recognize musical notes.	The experimental results showed that the algorithm greatly enhanced the speed of note recognition.

This paper extracted and recognized notes in the music based on the features of the music melody. Moreover, to enhance the recognition performance, the Mel-frequency cepstral coefficient (MFCC) features and music melody features were combined, and the convolutional neural network (CNN) algorithm was used as the classification recognizer. Finally, a self-created music library was used for simulation experiments. The novelty of this article lies in extracting melody features and MFCC features from music signals, then merging them into new music features, and using a CNN algorithm to recognize musical notes. This article provides an effective reference for accurately identifying musical notes in music signals.

2 Note recognition algorithm based on melody features

Melody is one of the core elements of music, and the style, genre, mood, and other characteristics of music can be identified through melody, and the music note sequence can also be extracted from the melody [7]. The above music characteristics can be used to realize music retrieval, matching, recommendation, and "duplicate detection" for copyright protection. The basic principle of traditional music note sequence extraction and identification is to extract the fundamental frequency of the pre-processed audio and then compare the fundamental frequency of the audio segment with the standard frequency of the note to identify the note sequence of the audio. However, in the actual situation, due to the interference of factors such as the environment, the timbre of the instrument itself, and the difference in audio rhythm, the fundamental frequency of the audio will have certain changes [8]. These changes may not involve the essential features but will also interfere with the classification and identification and impact the accuracy of note extraction and identification.

With the improvement of computer performance, the application fields of intelligent algorithms are also increasing, which also includes the note recognition of music melody [9]. The neural network algorithms are one type of intelligent algorithms. This kind of algorithm imitates the way of thinking of the human brain and fits the nonlinear law in the sample data by relying on the hidden layer in its structure. In this paper, the CNN algorithm is used as the note extraction and recognition classifier of the music melody [10]. The reason why CNN is selected as the classifier is that in the music note extraction and recognition, this paper not only uses the conventional audio features but also integrates the melody features to construct the fusion features of music. Due to the differences in the structure of the two features and the difficulty in determining the weight distribution ratio by using the weighted fusion method, the two features are superimposed [11] to form a feature combination similar to the matrix form, and the structure of the matrix form is just similar to the two-dimensional image. The process of note recognition based on melody features is described as follows.

① The audio data to be recognized is input.

② The audio is preprocessed by noise reduction and framing. The normalization processing is used to reduce the noise of the audio. The processing formula is:

$$\begin{cases} x' = x - \bar{x} \\ x'' = \dfrac{x'}{|x'|_{\max}} \end{cases}, \quad (1)$$

where x is the original signal, x' is the signal after eliminating direct current bias [12], $\bar{x}$ is the direct current bias component of the signal, and x'' is the signal after normalization. In framing process, the Hamming window function is sliding on the signal according to a specific step length, and the window function is employed to process the signal in the window each time to obtain a frame of processed audio signal. The related function is:

$$\begin{cases} S_w(n) = s(n) \times w(n) \\ w(n) = \begin{cases} 0 & else \\ 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \end{cases} \end{cases}, \quad (2)$$

where $S_w(n)$ is the audio signal after windowing, $s(n)$ is the signal after noise reduction processing, $w(n)$ is the window function, and N is the length of the signal.

③ Features such as MFCC features and melody features are extracted from the video signal [13], in which the pitch is used as the melody feature. Before extracting features, fast Fourier transform is conducted on the audio frame to obtain the spectrum of the audio frame. The extraction formulas of MFCC features and pitch melody features are:

$$\begin{cases} S(k)=\sum_{n=0}^{N-1}S_w(n)\cdot\exp\left(\frac{-2j\pi kn}{N}\right) \\ P(\omega)=|S(k)|^2 \\ P(m)=\ln\left(\sum_{k=0}^{N-1}P(\omega)\cdot H_m(k)\right) \\ c(l)=\sum_{m=1}^{M-1}P(m)\cos\left(\frac{\pi l(2m+1)}{2M}\right) \\ f_0=\arg\max\left\{\sum_{o=1}^{O}h_o A(of_i)\right\} \end{cases}, \quad (3)$$

where $S(k)$ is the spectrum signal obtained after fast Fourier transform [14], n is the time sampling point of the time domain signal, k is the sequence number of the sampling point, $P(\omega)$ is the instantaneous energy of $S(k)$, m is the sequence number of filters (M filters totally), $H_m(k)$ is the frequency response of the triangular filter, $c(l)$ is the L-order MFCC feature parameter, $P(m)$ is the energy spectrum function of the filtered signal in frequency domain, f_0 is the fundamental frequency calculated for the signal of the current frame, which represents the frequency with the highest level of confidence, O is the quantity of harmonics, h_o is the compression factor of the o-th harmonic, f_i is the i-th candidate fundamental frequency in $S(k)$, and $A(of_i)$ is the amplitude of the o-th harmonic when f_i is adopted as the fundamental frequency.

④ MFCC features and pitch melody features are combined into a rectangular structure, which is input into the CNN classifier for recognition. The CNN classifier includes input, convolution, pooling, and output layers, and the input and output layers are responsible for the input of feature data and note recognition results, respectively. As the hidden layer of CNN, the convolutional layer and the pooling layer extract the local and global features of the input data and compress them [14]. The convolutional layer is:

$$H_i=\sigma(H_{i-1}\otimes\omega_i+b_i), \quad (4)$$

where H_i and H_{i-1} are the feature maps output by the i-th and $i-1$-th layers, $\sigma(\cdot)$ is the activation function, b_i is the bias in the i-th layer structure, and ω_i is the weight in the layer structure. The pooling layer compresses the extracted local features, uses the pooling box to slide on the feature map, and merges the values in the box into one. There are two ways of merging: taking the mean and maximum [15].

⑤ The convolutional features after multiple convolutions and pooling are finally calculated in the fully connected layer of the CNN classifier to obtain the note recognition result.

3 Simulation experiments

3.1 Experimental data

A self-built note library was used for simulation experiments on the note recognition algorithm based on music melody. The simulation experiment was conducted on a laboratory server with the following configurations: Windows 11, 32 GB of memory, and an I7 processor. Music signals were collected in a soundproof room when building the note library. The instrument arrangement structure for collecting music signals is shown in Figure 2, including eight high-fidelity pickups (Speed Electron Company, China). The sampling rate was set to 48 kHz. The music signals of three instruments, piano, guitar, and recorder, were collected. The collected music signals were divided into two categories: one was a single-note music signal, while the other was a multi-note music signal.

For single-note music signals, each instrument was played according to the 25 notes shown in Table 1, and each note was played 30 times. For multi-note music signals, each instrument played 60 tracks (all 15 s in duration) composed of the notes shown in Table 2. There were 2,250 single-note music datasets in the sample set and 180 music tracks with a duration of 15 s in the multi-note music data set. 2/3 of each dataset was used as the training set, and the remaining was used as the test set.

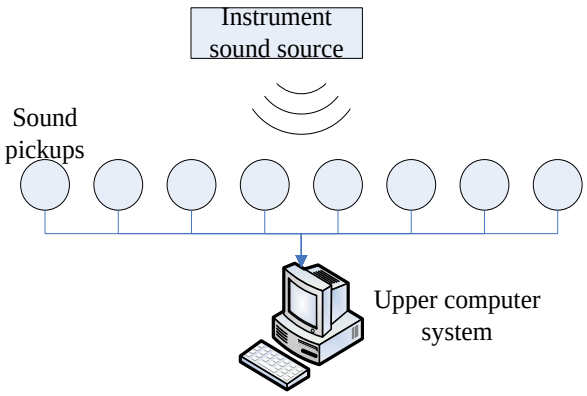

Figure 1: Arrangement structure of music signal acquisition.

Table 2: Types of notes used in simulation experiments.

Roll call	**Large group**	**Small group**	**1st small group**	**2nd small group**	**3rd small group**
1/do		#c	#c^1	#c^2	#c^3
2/re		d	d^1	d^2	d^3
3/mi		e	e^1	e^2	e^3
4/fa		#f	#f^1	#f^2	
5/so		g		g^2	
6/la	A	a	a^1	a^2	
9/si	B	b	b^1	b^2	

3.2 Experimental setup

Firstly, the performance of the note extraction and recognition algorithm under different MFCC feature dimensions and different CNN activation functions was tested to select the appropriate MFCC feature dimension and CNN activation function. The MFCC feature dimension was set to 6, 12, 24, 48, and 96, respectively. The activation function was set as rectified linear unit (ReLU), sigmoid, and tahn, respectively. The other parameters were set as shown in Table 3. The CNN classifier had three convolutional layers and three pooling layers.

Table 3: Relevant parameters of the CNN classifier.

Structure name	Parameter setting	Structure name	Parameter setting
Convolutional layer 1	32 convolutional kernels (2×1), a moving step length of 2	Pooling layer 1	2×1 pooling box, moving step length of 2, maximum pooling
Convolutional layer 2	64 convolutional kernels (2×1), a moving step length of 2	Pooling layer 2	2×1 pooling box, moving step length of 2, mean pooling
Convolutional layer 3	64 convolutional kernels (2×1), a moving step length of 2	Pooling Layer 3	2×1 pooling box, moving step length of 2, mean pooling
Learning rate	0.02	Training times	500

After determining the appropriate MFCC feature dimension and activation function, in order to verify the effectiveness of the proposed algorithm, a comparison experiment with the other two algorithms was carried out. The other two algorithms were the dynamic time warping (DTW)-based recognition algorithm and the CNN algorithm without music melody feature. The former was the algorithm mentioned above, which recognized notes by comparing the fundamental frequency of the audio with the standard frequency of the notes. The latter only used MFCC features as input, the rest were the same as the algorithm proposed in this paper, so the parameters on the CNN classifier were also the same. The indicators used in the comparison of the algorithms were precision, recall rate, and F-measure. Precision refers to the overall accuracy of an algorithm's recognition, while recall rate reflects the algorithm's positive detection rate, indicating the probability of correct identification in the recognition results. The F-measure is a comprehensive indicator that reflects the performance of an algorithm's recognition.

In addition, in order to demonstrate the degree of difference in performance among the three recognition algorithms, this study also employed independent t-tests to compare their recognition performance. When the p-value is less than 0.05, there is a significant difference between the two compared items.

3.3 Experimental results

Among the three indicators, precision, recall rate, and F-measure, F-measure is the comprehensive reflection of precision and recall rate. This indicator was used to compare the performance of the note extraction and recognition algorithm under different MFCC feature dimensions and activation functions, as shown in Table 4. With the increase of MFCC feature dimension, the performance of the algorithmfirst increased and then decreased under different activation functions. When the feature dimension was 24, the performance was the best. Under the same MFCC feature dimension, the performance of the algorithm using the sigmoid activation function was the best. Therefore, the MFCC feature dimension was finally set to 24, and the activation function of the CNN classifier was set to the sigmoid function.

Table 4: F-measure of the algorithm with different MFCC feature dimensions and activation functions.

MFCC feature dimension	6	12	24	48	96
Relu/%	82.1	88.7	92.3	83.4	75.8
Sigmoid/%	86.8	91.3	96.7	90.2	82.3
Tahn/%	80.7	87.4	90.8	81.2	73.4

Table 5 shows the note identification performance of the three algorithms for single-note music played by three instruments. It can be seen that the three algorithms showed different performances for single-note music played by different instruments, although the difference between the same algorithm was slightly smaller. The performance of the CNN algorithm based on music melody features was found to be the best when comparing the comprehensive performance of different algorithms for single-note extraction and recognition. The CNN algorithm without music melody feature was the second, and the algorithm based on DTW was the worst.

Table 5: Note recognition performance of algorithms for single-note music of three instruments.

Algorithm	Indicator	Piano	Guitar	Recorder	Synthesis
The DTW-	Precision /%	69.8	68.7	69.5	69.3

based recognition algorithm	Recall rate/%	68.9	67.9	67.8	68.2
	F-measure/%	69.3	68.3	68.6	68.8
The CNN algorithm without musical melody features	Precision/%	89.2 ●	88.9 ●	89.1●	89.1●
	Recall rate/%	88.7 ●	87.6 ●	8.3●	61.5 ●
	F-measure/%	88.9 ●	88.2 ●	15.2 ●	64.1 ●
The CNN algorithm based on music melody features	Precision/%	98.7 *+	97.6 *+	98.5*+	98.3 *+
	Recall rate/%	96.8 *+	96.9 *+	95.8*+	96.5 *+
	F-measure/%	97.7 *+	97.2 *+	97.1 *+	97.4 *+

Note: * indicates that the p value between the DTW algorithm and the proposed algorithm is less than 0.05; + indicates that the p value between the CNN algorithm and the proposed algorithm is less than 0.05; ● indicates that the p value between the DTW algorithm and the CNN algorithm is less than 0.05.

Table 6 shows the note recognition performance of the three recognition algorithms for multi-note music played by three instruments. It can be seen that for multi-note music played by different instruments, the three algorithms also showed different performances, and the same algorithm had little difference in music recognition performance played by different instruments. However, the CNN algorithm based on music melody features also had the best recognition performance among different algorithms. The CNN algorithm without music melody features was second, and the algorithm based on DTW was the worst. Moreover, by comparing the same algorithm between Table 5 and Table 6, it can be found that the recognition performance of the three algorithms decreased in the face of multi-note music.

Table 6: Note recognition performance of three recognition algorithms for multi-note music played by three instruments

Algorithm	Indicator	Piano	Guitar	Recorder	Synthesis
The DTW-based algorithm	Precision/%	42.1	43.5	43.6	43.1
	Recall rate/%	43.2	42.1	43.2	42.8
	F-measure/%	42.6	42.8	43.4	42.9
The CNN algorithm without musical melody features	Precision/%	75.4 ●	75.3 ●	76.4●	75.7 ●
	Recall rate/%	74.8 ●	74.6 ●	74.7●	74.7 ●
	F-measure/%	75.1 ●	74.9 ●	75.5●	75.2●
The CNN algorithm based on music melody features	Precision/%	92.4 *+	93.5 *+	93.1*+	93.0 *+
	Recall rate/%	91.7 *+	92.8 *+	92.9*+	92.5*+
	F-measure/%	92.0 *+	93.1 *+	93.0 *+	92.7*+

4 Discussion

With the rapid development of information technology and artificial intelligence, music recognition technology has become a hot research field. Among them, extracting and recognizing notes through melody features in music not only has important significance for music creation, teaching, and analysis but also holds great potential in areas such as music copyright protection and intelligent music recommendation. This article combined melody features in music signals with conventional MFCC features to form new music signal characteristics and then used a CNN algorithm to recognize notes in the music signal by taking advantage of complex convolution kernels that consider both local and global features. Afterward, simulation experiments were conducted. In the simulation experiments, the influence of MFCC feature dimension and types of activation functions in the CNN algorithm on algorithm recognition performance was tested first. Then, a comparative test was carried out between the proposed algorithm and two other algorithms in terms of single-note and multi-note aspects. When the MFCC dimension was 24 and the sigmoid activation function was used in the CNN algorithm, the algorithm achieved the best recognition performance. This is because when the MFCC feature dimension is too small, it cannot reflect the characteristics of music well. If it is too large, it will cause redundancy without significant improvement in feature effectiveness but greatly increase computational complexity. The sigmoid activation function can better fit nonlinear patterns compared to the other two activation functions.

The comparison of the proposed algorithm with two other recognition algorithms showed that the proposed algorithm had better performance for both single and multiple notes. The reasons for the above results were analyzed. The recognition algorithm based on DTW extracted and recognized notes by comparing the fundamental frequency of the music with the standard fundamental frequency of the note, which was relatively simple in principle. Moreover, the music segment and the

standard segment were aligned by way of DTW, but it still cannot fully fit the nonlinear law. In the CNN algorithm without music melody features, the convolutional layer was used to extract the local features in the audio signal and form the global features, which effectively fit the nonlinear law, so the recognition performance was higher. The CNN algorithm based on music melody features combined the music melody features and the basic audio features to further expand the features of music, so the recognition performance was further improved.

The novelty of this article lies in extracting melody features and MFCC features from music signals, then combining them into new music features, and using the CNN algorithm to recognize notes. This article provides an effective reference for accurately identifying notes in music signals.

5 Conclusion

This paper extracted and recognized the notes in the music based on the features of the music melody. Moreover, to improve the recognition performance, the music melody features were combined with MFCC features, and the CNN algorithm was used as the classification recognizer. Experiments were performed using a single-note and multi-note music library constructed by three instruments: piano, guitar, and recorder. The performance of the proposed algorithm under different MFCC feature dimensions and CNN activation functions was tested, and then it was compared with the DTW algorithm and the CNN algorithm without music melody features. The following results were obtained. (1) When the MFCC feature dimension was set to 24 and the activation function of the CNN classifier was set to sigmoid, the performance of the proposed algorithm was the best. (2) For single notes, the recognition performance of the CNN algorithm based on music melody features was the best, followed by the CNN algorithm without music melody features, and the algorithm based on DTW was the worst. (3) For multiple notes, the recognition performance of the proposed algorithm was still the best, but compared with the single note, the recognition performance of the three algorithms was reduced.

The limitation of this article lies in the improvement only on the characteristics of music signals and the limited generalization of the algorithm due to a small number of training samples. Therefore, future research directions include improving the recognition classifier and increasing the sample size to enhance the generalization of the recognition algorithm.

6 References

[1] Xiao Z, Chen X, Zhou L (2019). Real-time optical music recognition system for dulcimer musical robot. *Journal of Advanced Computatioanl Intelligence and Intelligent Informatics*, 23(4), pp. 782-790. https://doi.org/10.20965/jaciii.2019.p0782.

[2] Cui J, Wang H (2022). Multimedia music image production combining HMM and song feature tags. *Journal of Electronic Imaging*, 31(5), pp. 51419.1-51419.15.

[3] Liu L, Leung MF (2022). Lute acoustic quality evaluation and note recognition based on the softmax regression bp neural network. *Mathematical Problems in Engineering: Theory, Methods and Applications*, 2022(Pt.13), pp. 1.1-1.7. https://doi.org/10.1155/2022/1978746.

[4] Tamboli AI, Kokate RD (2017). An effective optimization-based neural network for musical note recognition. *Journal of Intelligent Systems*, 28(1), pp. 173-183. https://doi.org/10.1515/jisys-2017-0038.

[5] Cinar GT, Sequeira PMN, Principe JC (2017). Hierarchical linear dynamical systems for unsupervised musical note recognition. *Journal of the Franklin Institute*, 355(4), pp. 1638-1662. https://doi.org/10.1016/j.jfranklin.2017.04.013.

[6] Nazar A, Ramo FM (2021). A new parallel bat algorithm for musical note recognition. *International Journal of Electrical and Computer Engineering*, 11(1), pp. 558-566. https://doi.org/10.11591/ijece.v11i1.pp558-566.

[7] Herff SA, Olsen KN, Dean RT (2018). Resilient memory for melodies: The number of intervening melodies does not influence novel melody recognition. *The Quarterly Journal of Experimental Psychology Section B*, 71(5), pp. 1150-1171. https://doi.org/10.1080/17470218.2017.1318932.

[8] Bomgardner M (2021). MATERIALS Schlumberger pilots new lithium extraction. *Chemical and Engineering News: "news edition" of the American Chemical Society*, 99(11), pp. 10.

[9] Wang Y (2021). Research on handwritten note recognition in digital music classroom based on deep learning. *Journal of Internet Technology*, 2021(6), pp. 22.

[10] Li X, Robin H (2019). Construction and analysis of hidden Markov model for piano notes recognition algorithm. *Journal of Computer Science Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 37(3 Pt.1), pp. 3293-3302.

[11] Wu R (2021). Research on automatic recognition algorithm of piano music based on convolution neural network. *Journal of Physics: Conference Series*, 1941(1), pp. 1-7. https://doi.org/10.1088/1742-6596/1941/1/012086.

[12] Xiao S, Hu Y, Han J, Zhou R, Wen J (2016). Bayesian networks-based association rules and knowledge reuse in maintenance decision-making of industrial product-service systems. *Procedia CIRP*, 47, pp. 198-203. 10.1016/j.procir.2016.03.046.

[13] Rao W, Zhu L, Pan S, Yang P, Qiao J (2019). Bayesian network and association rules-based transformer oil temperature prediction. *Journal of Physics Conference*, 1314(1), pp. 1-8. https://doi.org/1314:012066.

[14] Siddiquee MR, Rahman S, Chowdhury SUI, Rahman MR (2016). Association rule mining and audio signal processing for music discovery and recommendation. *International Journal of Software Innovation*, 4(2), pp. 71-87. https://doi.org/10.4018/IJSI.2016040105

[15] Lai WH, Lee CY (2016). Query by singing/humming system using segment-based melody matching for music retrieval. *WSEAS Transactions on Systems*, 15, pp. 157-167.