AUTOMATIC CHINESE NATIONAL PENTATONIC MODES RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Zhaowen Wang[1] **Mingjin Che**[2] **Yue Yang**[1] **Wenwu Meng**[2]
Qinyu Li[2] **Fan Xia**[2] **Wei Li**[3,4]

[1] Department of Music AI and Information Technology, Central Conservatory of Music, China
[2] College of Experimental Art, Sichuan Conservatory of Music, China
[3] School of Computer Science and Technology, Fudan University, China
[4] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China

{wzw, yewyang}@mail.ccom.edu.cn, weili-fudan@fudan.edu.cn,
{993393523, 1056844091, 935362804, 409769992}@qq.com

ABSTRACT

Chinese national pentatonic modes, with five tones of Gong, Shang, Jue, Zhi and Yu as the core, play an essential role in traditional Chinese music culture. After the early twentieth century, with the development of new Chinese music, the ancient Chinese theory of scales gradually developed into a new pentatonic modes theory under the influence of western music. In this paper, we briefly introduce our self-built CNPM (Chinese National Pentatonic Modes) Dataset, then design residual convolutional neural network models to identify which TongGong system the mode belongs, the pitch of tonic, the mode pattern and the mode type from audio signals, in combination with musical domain knowledge. We use both single-task and multi-task models with three strategies for identification, and compare them with a simple template-based baseline method. In experiments, we use seven accuracy metrics to evaluate the models. The results on identifying both the tonic pitch and the pattern of mode correctly achieve an average accuracy of 69.65%. As an initial research on automatic Chinese national pentatonic modes recognition, this work will contribute to the development of multicultural music information retrieval, computational ethnomusicology and five-tone music therapy.

1. INTRODUCTION

The modern Chinese national pentatonic modes theory based on the five tones of Gong, Shang, Jue, Zhi and Yu was created by Chinese musicians who combined music theories such as absolute pitch, twelve-tone equal temperament and major/minor modes used in western music with Chinese music theory after the early twentieth century.

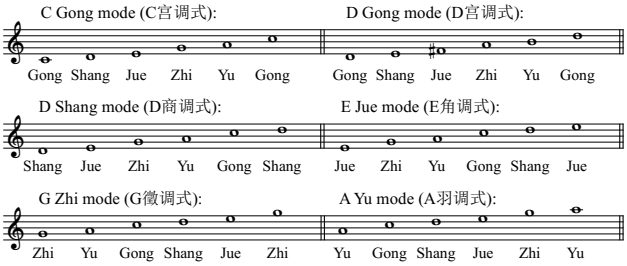

Figure 1. Examples of mode scales. Chinese national pentatonic modes include pentatonic mode, hexatonic mode and heptatonic mode which are based on the five tones. Only pentatonic mode scales are shown here. The C Gong mode and the D Gong mode share the same mode pattern but with different pitch of tonic. The C Gong mode and the four modes in the second and third rows all belong to the C TongGong system.

From the 1920s to the 1960s, Chinese musicians such as Guangqi Wang [1], Zhengya Wang [2] and Yinghai Li [3] gradually proposed the theory of Chinese national pentatonic modes based on the ancient Chinese scales and related concepts, with reference to western scales and modes. The theory was later written by Chongguang Li into the basic music theory textbook [4], and became the version commonly used in Chinese music teaching. The musical data selected and the analyses conducted in this paper are based on the modern Chinese national pentatonic modes theory. It can be applied to most of the traditional music pieces, especially those works of Chinese national orchestral instruments adapted from traditional tunes. Further explanation and clarification of the theory and ancient Chinese music are listed here [1] .

Gong, Shang, Jue, Zhi and Yu can be called step names in the modern theory. They have a fixed interval relationship which can be described as Gong-Shang major second, Shang-Jue major second, Jue-Zhi minor third, Zhi-Yu major second and Yu-Gong minor third. In addition, Qingjue indicates the tone a minor second above Jue. Biangong and

[1] https://github.com/Hellwz/CNPM-ISMIR2022/blob/main/notes.md

Type	Step names
Pentatonic	Gong / C - Shang / D - Jue / E - Zhi / G - Yu / A
Hexatonic (Qingjue)	Gong / C - Shang / D - Jue / E - Qingjue / F - Zhi / G - Yu / A
Hexatonic (Biangong)	Gong / C - Shang / D - Jue / E - Zhi / G - Yu / A - Biangong / B
Heptatonic Yayue	Gong / C - Shang / D - Jue / E - Bianzhi / $^{\sharp}$F - Zhi / G - Yu / A - Biangong / B
Heptatonic Qingyue	Gong / C - Shang / D - Jue / E - Qingjue / F - Zhi / G - Yu / A - Biangong / B
Heptatonic Yanyue	Gong / C - Shang / D - Jue / E - Qingjue / F - Zhi / G - Yu / A - Run / ^{b}B

Table 1. Six mode types in Chinese national pentatonic modes. The pitch after each step name is just one of the possible matches, listed for better understanding.

Bianzhi indicate the tones a minor second below Gong and Zhi, respectively. And Run [2] is the tone a major second below Gong. The original Gong, Shang, Jue, Zhi and Yu tones are called Zheng-tones (i.e. main tones), and others are considered as Bian-tones (i.e. changed tones). Figure 1 shows several pentatonic scales for better understanding.

A complete mode name in Chinese national pentatonic modes can be divided into three parts: the pitch of tonic, the mode pattern, and the mode type. The pitch of tonic is indicated by pitch names such as C, D and E. The mode pattern has five categories according to the step name of tonic. Gong mode use Gong tone as tonic, Shang mode use Shang tone as tonic, and by analogy, there are five mode patterns of the Gong, Shang, Jue, Zhi and Yu. The mode type has six categories which are called Pentatonic, Hexatonic (Qingjue), Hexatonic (Biangong), Heptatonic Yayue, Heptatonic Qingyue and Heptatonic Yanyue mode. The step names these six mode types contain are shown in Table 1. Any Zheng-tone (i.e. Gong, Shang, Jue, Zhi and Yu) in it can be used as the tonic to form a mode. Therefore, combining the pitch of tonic, we get a total of 12x5x6=360 kinds of modes theoretically.

There is also a unique concept in Chinese national pentatonic modes theory called the TongGong system. If the Gong's pitch of two modes are the same, they are said to belong to the same TongGong system. For instance, the C Gong mode, D Shang mode, E Jue mode, G Zhi mode and A Yu mode all belong to the C TongGong system. The TongGong system to which the mode belongs, the tonic pitch and the pattern of the mode are related. Identifying any two of them can infer the third one. This is also a main idea of the approach in this paper.

Automatic key/mode recognition is an important research direction in MIR, but little research has been done on automatic Chinese national pentatonic modes recognition for audio. The main reasons include the lack of annotated data, the variety of national modes, and the difficulty of accurately identifying the pitch of Chinese instruments, etc. These reasons make recognition more difficult.

In this paper we develop deep learning models based on residual convolutional neural network for automatic recognition of Chinese national pentatonic modes. By entering the spectrogram of the audio, the models will output the full name of the national mode it belongs to. The correct recognition of the mode pattern and the pitch of tonic is the main focus of this paper. The benefits of using neural networks are that they can automatically extract features from spectrograms and have better generalizability to different musical genres and instruments. We use both single-task and multi-task models with three strategies, as described in section 3, for identification, and compared them with a simple template matching approach, as shown in section 4. Data augmentation is also adopted for better results.

The significance of Chinese national pentatonic modes automatic recognition includes: 1) Strengthening computer's understanding of music that can assist humans in musicological and ethnomusicological analysis; 2) Used to carry out the research of five-tone music therapy. Music in Chinese national pentatonic modes can regulate the body, assist in treatment and has a positive effect on some mental diseases [5, 6]; 3) Enhancing and further understanding the algorithmic mechanism of automatic mode recognition itself; 4) Assisting in the preservation of traditional music culture; 5) Auxiliary to other MIR studies.

2. RELATED WORK

The related work on automatic mode recognition and the application of convolutional neural networks in Chinese music information retrieval are introduced in this section.

2.1 Automatic Key/Mode Recognition

Automatic Key/Mode Recognition is one of the core tasks of MIR. The early studies on western major and minor key recognition were traditionally handled by template matching and Hidden Markov Models (HMMs). [7] proposed 24 major/minor key templates and algorithms for automatic key detection. [8] improved these templates and focused more on harmonic minor in minor key templates. [9] calculated Harmonic Pitch Class Profile (HPCP) for key matching of audio. [10] obtained templates from real instruments. [11] utilized probabilistic models (HMMs), combined with templates, to assist in key recognition. [12] extracted Pitch Class Profile (PCP) sequences and then used a single HMM directly.

[2] Run is also referred to as Qingyu (a minor second above Yu) by some scholars.

Classification basis	Names	Amount	Labels
TongGong System	C; ♯C (^{b}D); D; ♯D (^{b}E); E; F; ♯F (^{b}G); G; ♯G (^{b}A); A; ♯A (^{b}B); B	12	0-11
Pitch of Tonic	C; ♯C (^{b}D); D; ♯D (^{b}E); E; F; ♯F (^{b}G); G; ♯G (^{b}A); A; ♯A (^{b}B); B	12	0-11
Mode Pattern	Gong; Shang; Jue; Zhi; Yu	5	0-4
Mode Type	Pentatonic; Hexatonic (Qingjue); Hexatonic (Biangong); Heptatonic Yayue; Heptatonic Qingyue; Heptatonic Yanyue	6	0-5

Table 2. Definition of the mode category

With the development of deep neural networks, Convolutional Neural Network (CNN) has also been applied to the study of automatic key recognition. [13] proposed an end-to-end framework based on CNN for global key detection and can be adapted to different music styles. [14] improved the above model by removing the dense layer and keeping only the convolutional and pooling layers to obtain a genre-independent model and enhance generalization ability. [15] recognized local key for classical music, comparing the HMM and CNN methods and the effect of segmenting the dataset by song or version. [16] compared the influence of different directional convolutional architectures on the results along VGG-style networks. [17] designed a complex CNN architecture based on InceptionV3 to recognize 24 major/minor keys and evaluated a broad range of data augmentation methods.

In the literature, we found three works related to the recognition of Chinese national pentatonic modes, with two from symbolic format and only one from audio signals. None of them used neural networks. [18] used a manually designed decision tree for Chinese pentatonic mode recognition, and [19] used a template-matching based algorithm for automatic recognition of Chinese pentatonic mode and heptatonic mode. These two studies aim at MIDI files that are more conducive to recognizing musical meta-information such as pitch rather than audio signals. [20] designed uniform and ordinal templates for mode recognition of guqin music, using template matching to identify Gong, Shang and Zhi pentatonic modes from audio. We compare this method with our baseline in subsection 4.2.

2.2 Convolutional Neural Networks in CMIR

There have been several studies related to Chinese national music in the field of MIR in recent years, and some relevant datasets have emerged [21–24]. CNN models such as VGG [25], GoogLeNet [26] and ResNet [27] have made great progress on image recognition tasks. Inspired by this, CNN has also been applied in the CMIR (Chinese Music Information Retrieval) field. [28] utilized Convolutional Recurrent Neural Network (CRNN) for activity detection of Chinese national polyphony instruments and achieved a temporal resolution of seconds. [29] extracted Mel Frequency Cepstral Coefficients (MFCC) features of the audio and used VGGish network to classify 78 Chinese musical instrument timbres. [30] adopted Fully Convolutional Networks (FCN) to achieve the best results in the playing technique detection task of Chinese Erhu and Bamboo Flute instruments.

3. METHODS

The methods we propose and use are described below, including template matching and neural network models.

3.1 Definition

According to the definition in this paper, there are 360 kinds of Chinese national modes theoretically. We show their category definitions based on different classification basis and the numbering method of data annotation in Table 2. For simplicity, in the following we refer TongGong system to which the mode belongs as "system", pitch of the tonic as "tonic", mode pattern as "pattern", and mode type as "type". The primary task when classifying is to identify pattern and tonic, with system as auxiliary item, followed by type classification as a secondary task.

Based on the tonic t and the system s, we can infer the mode pattern. When t equals s, it is Gong mode. When t is 2 semitones higher than s, it is Shang mode. 4 semitones higher is Jue mode, 7 semitones higher is Zhi mode and 9 semitones higher is Yu mode.

3.2 Baseline

In order to compare the knowledge-based method with the data-driven method (i.e. neural network method), we use a template matching method designed for Chinese national pentatonic modes as the baseline.

The chroma features of the entire audio are obtained using the librosa package [31] and summed to get a twelve-dimensional chroma vector which reflects the energy of each pitch among the whole audio without octave information. Firstly, classify the TongGong system to which the audio belongs. The template of C TongGong system is [1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0] and others can be obtained by shifting the template in a loop. The Pearson correlation coefficient between chroma vector and each template is calculated respectively. The larger the correlation coefficient is, the higher the matching degree is. The template with the maximum value is the recognition result. Then for the tonic pitch, since most of the music we analyze returns to the tonic at the end, we use a simple way to identify the tonic: directly sum the chroma features of the last 500 frames according to the pitches, and regard the pitch name

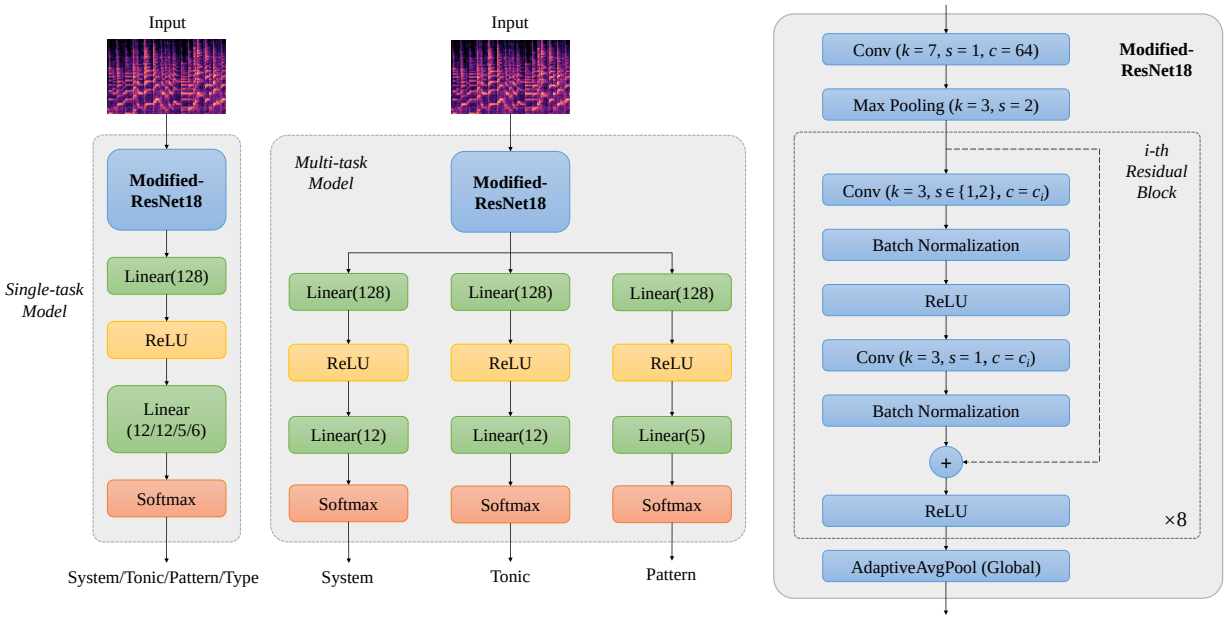

Figure 2. Model architectures. The shape of input is [Batch_size, Channel=1, F=168, T]. There are 4 single-task models for predicting TongGong system, tonic pitch, mode pattern and type respectively. Specifically, the single-task model for tonic prediction has no linear layer of 128 neurons and the following ReLU layer. In modified-ResNet18, k means the kernel size is $k \times k$, s indicates the stride and c stands for the number of channels. For 3rd, 5th and 7th residual block, the first s of conv layer equals 2, and others are 1. $c_i \in$ [64, 64, 128, 128, 256, 256, 512, 512]. A 1×1 convolutional layer is also used on the dashed line in the residual block if the number of channels does not match.

corresponding to the maximum value as the tonic pitch. As for the mode type recognition, the templates consisting of 0s and 1s are obtained according to the scale corresponding to each mode, and the results are acquired using a calculation method similar to that of the TongGong system identification. At last, the pattern of the mode can be obtained through the TongGong system and the tonic pitch.

3.3 CNN Models

Our CNN models use a modified ResNet18 [27] structure as the backbone. ResNet models have reached SOTA on multiple classification tasks in history and have simple structures which make them easy to tune and train. Their residual blocks are also widely used in various neural networks. For the input of our models, we use Constant-Q Transform (CQT) spectrogram, which is more appropriate for music audio than other spectrograms and contains more information than chroma features. Referring to the configuration of [14] and [17], we set the spectrogram frequency range from C1 to C8 with 168 frequency bins totally, and then downsample to 5 frames per second in the time axis. Because of the specific structure of our models, we can use any length of spectrogram as input.

The architectures of our models are shown in Figure 2. Compared to the original ResNet, our main changes include: 1) Modify the three channels of the input to a single channel for audio task; 2) Remove the downsampling from the first convolutional layer to accommodate a smaller size input; 3) Add the fully-connected layer in the output section to facilitate establishing different mapping relationships for different subtasks, and fit our tasks of less categories, compared to image classification task.

Due to the diverse types of modes but small amount of data, it is not reasonable to directly classify 60 or 360 categories. Therefore we use the following three strategies for recognition: 1) Directly use two single-task models to predict the tonic and pattern respectively; 2) Use two single-task models to predict the system and tonic respectively, and then indirectly calculate the pattern from these two results; 3) Use one multi-task model to recognize system, tonic and pattern, where pattern can be derived both directly and indirectly. When training the multi-task model, the losses of three outputs are summed for back propagation. As for type recognition, due to its unbalanced data distribution, difficulty of identification and little relationship with the other three, we directly use a single model to predict without adding it to multi-task model.

For the training input, we divide the spectrogram without overlap into 20s snippets, which is the same length as [17], to perform mini-batch training. Especially, the last snippet must be taken out regardless of whether it overlaps with the previous one, because it is critical for the tonic recognition. All snippets in the training set will be involved in the training, except that only the last snippet of each recording will be used when training the tonic model due to the characteristic of music. The tonic labels of non-final snippets will be masked when training the multi-task model for the same purpose, with other labels unchanged.

In the inference process, it is possible to either input the whole spectrogram directly or divide it into snippets,

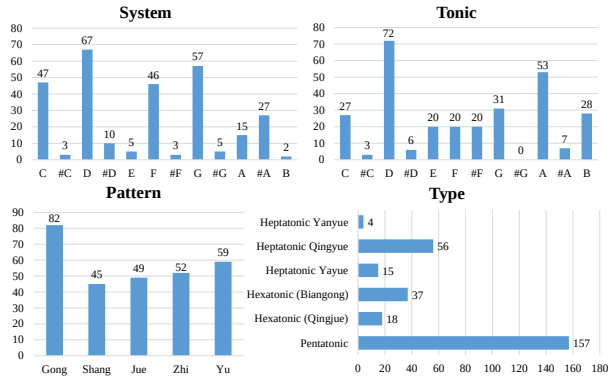

Figure 3. Data distribution of the CNPM Dataset

and then combine the results of each snippet to obtain the final result. In preliminary experiments we discover that for system, pattern and type tasks, using the whole spectrogram as input gives better results. In contrast, the tonic recognition only requires the last snippet of each song.

4. EXPERIMENTS

In the following we present the dataset we use and the experimental results we obtain using different methods.

4.1 Dataset

We use and expand our self-built CNPM (Chinese National Pentatonic Modes) Dataset. After expanding, the dataset contains 287 music recordings, including instrumental and vocal music, with instruments mainly being traditional Chinese instruments such as Guzheng, Guqin, Pipa, etc. The complete mode information is also included as labels, annotated by faculty and students in related disciplines. The average audio duration is 179.5s and the distribution of data is shown in Figure 3. Most of the audio comes from the Internet, with some Shang and Jue mode pieces being created and recorded by our musicians. In collecting and labeling the data, we select music pieces with clearer modes, more stable pitches, and mostly performed by contemporary musicians to avoid controversy in judging the modes. For music with modulation, we divide and label them accordingly. The dataset can be found here [3] .

The distribution of mode pattern in the database is relatively balanced, but system and tonic are not. To address the problem of fewer samples in some categories, we perform data augmentation by pitch shifting. To be specific, our processing of audio includes ascending a semitone, ascending a whole tone, descending a semitone and descending a whole tone with labels changing accordingly. Note that this augmentation only changes system and tonic, not pattern or type. And it is only applied to the training data.

4.2 Results

To ensure the effectiveness of the baseline approach, we first compare it with the 12-D ordinal template matching method proposed in [20] which also uses the chroma feature of audio. Different from our indirect prediction of five mode patterns from system and tonic, the ordinal templates directly predict three mode patterns of Gong, Shang and Zhi. The recognition accuracy is shown in Table 3, and our baseline method achieves better results.

For the neural network training, we use the Adam optimizer, cross-entropy loss function and a learning rate of 0.001. The batch size of each model is set to 32. For recognition results, we develop seven accuracy metrics to evaluate, as shown in Table 4.

In preliminary experiments, we adopt and modify the main structure of the InceptionKeyNet [17] network which is designed for major/minor keys recognition to perform our modes recognition. The usage method is similar to the modified-ResNet single-task models. We conduct experiments on the same random test set and find the two models have similar performance. But because the original InceptionKeyNet is implemented by TensorFlow and our models are implemented by PyTorch, it is hard to achieve a completely fair comparison and claim which structure is better. Since our models based on ResNet have a simpler structure and can converge in fewer epochs, making them easier to implement and train, the following experiments are performed using modified-ResNet18-based models.

The main experiments are conducted on the template-based baseline approach, the ResNet-based single-task models, as well as the ResNet-based multi-task model, and the results are shown in Table 5. To fully evaluate the models, we use a 5-fold cross-validation to obtain more con-

[3] https://ccmusic-database.github.io/en/database/ccm.html

Accuracy (%)	Ordinal	Baseline
Gong	62.20	**71.95**
Shang	55.56	**64.44**
Jue	-	**63.27**
Zhi	46.15	**67.31**
Yu	-	**62.71**

Table 3. Accuracy results of mode pattern recognition using template matching methods.

Metric	Object
ACC1	System
ACC2	Tonic
$ACC3_D$	Pattern (Directly)
$ACC3_I$	Pattern (Indirectly)
ACC4	Tonic and Pattern
ACC5	Type
ACC6	Tonic, Pattern and Type

Table 4. Accuracy metrics. Indirectly here means we get the mode pattern via system and tonic recognition results. Higher ACC4 is our primary task. And ACC6 means all correct.

ACC (%)	Baseline	Res-ST	Res-MT
ACC1	85.71	**88.50±1.80**	85.36±1.44
ACC2	71.78	74.85±9.28	**79.07±3.25**
$ACC3_D$	-	58.58±6.24	**63.79±7.74**
$ACC3_I$	66.55	71.01±10.30	**71.74±5.64**
ACC4	64.46	69.27±9.70	**69.65±5.21**
ACC5	48.08	**57.87±4.81**	-
ACC6	31.01	**42.15±3.65**	-

Table 5. Accuracy results under 5-fold cross-validation. Baseline is the template matching method. Res-ST is short for the modified-ResNet18 single-task model and Res-MT is the modified-ResNet18 multi-task model.

vincing accuracy results and compare them with the baseline method acting on the whole dataset. When dividing the data, we divide it by music tracks, keep the mode pattern ratio constant and ensure that only the corresponding training data are involved in back propagation in each fold of training.

As seen in the results, CNN-based methods outperform template matching method, and indirect mode pattern prediction is much better than direct prediction. The correct system and tonic predictions can introduce the correct pattern, but the correct pattern prediction can also be brought out by the wrong but relatively correct system and tonic. Therefore we need ACC4 to make sure all three (i.e. system, tonic and pattern) are correct. Since $ACC3_I$ is higher, we use the pattern results of indirect prediction to calculate ACC4 and ACC6. The accuracy of the system, tonic and pattern being predicted correctly at the same time achieved by our ResNet-based models is a satisfactory result. Due to the very uneven data distribution of mode type, the model responsible for type prediction is hard to train. So ACC5 and ACC6 are listed for reference only, and are not the primary tasks of this paper. In the comparison of single-task model and multi-task model, we find that the multi-task model has better ability to identify the tonic and directly recognize the mode pattern, which indicates the effectiveness of the multi-task learning strategy. Then for $ACC3_I$ and ACC4, the single-task and multi-task models perform similarly. The multi-task model performs more consistently over different folds with less parameters compared to using multiple single-task models.

Due to the relatively good performance of the multi-task model, we combined the indirect mode pattern recognition results of each fold to obtain a normalized confusion matrix and visualized it, as shown in Figure 4. Note that if the result of a certain indirect prediction is invalid, we use the direct prediction instead. From the confusion matrix, we can see that the CNN model has a high recognition ability for Gong, Jue and Yu modes, but the recognition of Shang and Zhi modes needs to be improved. This difference should be more related to the quality, distribution and representativeness of the data than the model itself. Besides, a larger value of k in k-fold cross-validation may lead to better results.

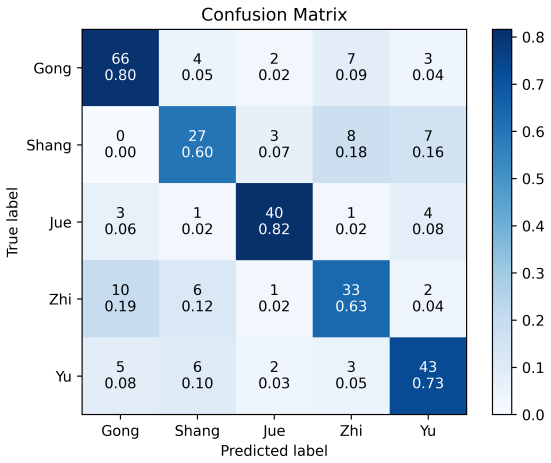

Figure 4. Normalized confusion matrix for mode pattern recognition results using the multi-task CNN model.

5. CONCLUSION & FUTURE WORK

In this paper, we explore the automatic mode recognition for Chinese national music composed after the early twentieth century with the characteristics of Chinese national pentatonic modes. For music from diverse ethnic and cultural backgrounds, the musical concepts and logic will be different, as well as the use of different tuning systems, rhythmic patterns, instruments and expressions. Traditional Chinese music played by traditional Chinese instruments is rich in the use of tuning systems, including compound tuning systems. This requires researchers to design more appropriate algorithms and models based on the characteristics of the music in order to get better results. Meanwhile, this can also facilitate the development of mainstream MIR algorithms.

We combine the musical domain knowledge when designing the methodology used in this paper. When identifying the mode pattern, we do so indirectly by considering its relationship to the TongGong system and the tonic pitch. As for the pitch of tonic, only the ending fragment of the music is considered according to the musical characteristic. We also use multi-task learning for recognition based on the correlation between system, tonic and pattern.

The accuracy of mode type is relatively low due to the difficulty of recognition and uneven data. In the future, more detailed annotation with segmentation can be carried out for audio to improve the accuracy. It is also possible to use other methods to identify each instrument and its pitches first before recognizing, but this remains a challenge for traditional Chinese instruments.

For the issue of small data size, more data augmentation methods can be used, such as time scaling, random masking, adding noise, etc. Transfer learning can also be utilized to train the backbone neural network on a large dataset first to have the ability of capturing tonality and pitch features, and then on a small dataset to accomplish a specific task.

6. ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China (2019YFC1711800), NSFC (62171138). Wei Li and Fan Xia are co-corresponding authors of this paper.

7. REFERENCES

[1] G. Wang, *Music of the oriental nation (in Chinese)*. Shanghai, China: Zhong Hua Book Company, 1929.

[2] Z. Wang, *The pentatonic scale and its harmony (in Chinese)*. Shanghai, China: Wenguang Bookstore, 1949.

[3] Y. Li, *The mode and harmony of Han nationality (in Chinese)*. Shanghai, China: Shanghai Literature and Art Publishing House, 1959.

[4] C. Li, *Fundamentals of music theory (in Chinese)*. Beijing, China: Music Publishing House (now People's Music Publishing House), 1962.

[5] J. Zhang, F. Xu, J. Du, X. Tan, C. Wu, and J. Kong, "Exploring and analyzing the five tone therapy in Chinese medicine (in Chinese)," *Journal of Changchun University of Traditional Chinese Medicine*, vol. 27, no. 5, pp. 702–704, 2011.

[6] S. Mi and M. Shi, "Discussion on the therapy of Wuyin (in Chinese)," *Clinical Journal of Chinese Medicine*, vol. 11, no. 29, pp. 12–14, 2019.

[7] C. L. Krumhansl, *Cognitive foundations of musical pitch*. Oxford, UK: Oxford University Press, 1990.

[8] D. Temperley, "What's key for key? the krumhansl-schmuckler key-finding algorithm reconsidered," *Music Perception*, vol. 17, no. 1, pp. 65–100, 1999.

[9] E. Gómez and P. Herrera, "Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[10] Ö. Izmirli, "Template based key finding from audio," in *Proceedings of the 2005 International Computer Music Conference (ICMC)*, Barcelona, Spain, September 2005, pp. 211–214.

[11] G. Peeters, "Musical key estimation of audio signal based on hidden markov modeling of chroma vectors," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, September 2006, pp. 127–131.

[12] W. Chai and B. Vercoe, "Detection of key change in classical piano music," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005, pp. 468–473.

[13] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, August 2017, pp. 966–970.

[14] F. Korzeniowski and G. Widmer, "Genre-agnostic key classification with convolutional neural networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, September 2018, pp. 264–270.

[15] C. Weiß, H. Schreiber, and M. Müller, "Local key estimation in music recordings: A case study across songs, versions, and annotators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 2919–2932, 2020.

[16] H. Schreiber and M. Müller, "Musical tempo and key estimation using convolutional neural networks with directional filters," in *Proceedings of the 16th Sound & Music Computing Conference (SMC)*, Málaga, Spain, May 2019, pp. 47–54.

[17] S. A. Baumann, "Deeper convolutional neural networks and broad augmentation policies improve performance in musical key estimation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, Online, November 2021, pp. 42–49.

[18] Y. Deng, L. Zhou, S. Ni, S. Zhang, and M. You, "Research on the recognition algorithm of Chinese traditional scales based on decision tree," in *37th Chinese Control Conference (CCC)*, Wuhan, China, July 2018, pp. 9527–9534.

[19] M. You, L. Chen, L. Zhou, and J. He, "Research on Chinese national music mode recognition based on template matching (in Chinese)," *Journal of Fudan University (Natural Science)*, vol. 59, no. 3, pp. 262–269, 2020.

[20] Y.-F. Huang, J.-I. Liang, I.-C. Wei, and L. Su, "Joint analysis of mode and playing technique in guqin performance with machine learning," in *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, October 2020, pp. 85–92.

[21] Z. Li, S. Yu, C. Xiao, Y. Geng, W. Qian, Y. Gao, and W. Li, "CCMusic Database: Construction of Chinese music database for MIR reaserch (in Chinese)," *Journal of Fudan University (Natural Science)*, vol. 58, no. 3, pp. 351–357, 2019.

[22] Z. Li and B. Han, "On database construction of acoustic system of Chinese traditional instrumental (in Chinese)," *Chinese Musicology*, no. 2, pp. 92–102, 2020.

[23] X. Liang, Z. Li, J. Liu, W. Li, J. Zhu, and B. Han, "Constructing a multimedia Chinese musical instrument database," in *Proceedings of the 6th Conference on Sound and Music Technology (CSMT)*, ser. Lecture Notes in Electrical Engineering (LNEE), vol. 568. Springer, 2019, pp. 53–60.

[24] X. Gong, Y. Zhu, H. Zhu, and H. Wei, "ChMusic: A traditional Chinese music dataset for evaluation of instrument recognition," in *4th International Conference on Big Data Technologies (ICBDT)*, Qingdao, China, September 2021, pp. 184–189.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, May 2015.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, June 2015, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, June 2016, pp. 770–778.

[28] Z. Li, C. Jiang, X. Chen, Y. Ma, and B. Han, "Instrument activity detection of China national polyphonic music based on convolutional recurrent neural network (in Chinese)," *Journal of Fudan University (Natural Science)*, vol. 59, no. 5, pp. 511–516, 2020.

[29] R. Li, Y. Xie, Z. Li, and X. Li, "Chinese musical instruments classification using convolutional neural network (in Chinese)," *Journal of Fudan University (Natural Science)*, vol. 59, no. 5, pp. 517–522, 2020.

[30] Z. Wang, J. Li, X. Chen, Z. Li, S. Zhang, B. Han, and D. Yang, "Deep learning vs. traditional MIR: a case study on musical instrument playing technique detection," in *Proceedings of 13th International Workshop on Machine Learning and Music (MML) at ECML/PKDD*, September 2020, pp. 5–9.

[31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference (SciPy)*, Austin, USA, July 2015, pp. 18–25.