Classification basis	**Names**	**Amount**	**Labels**
TongGong System	C; $^\sharp$C (^{b}D); D; $^\sharp$D (^{b}E); E; F; $^\sharp$F (^{b}G); G; $^\sharp$G (^{b}A); A; $^\sharp$A (^{b}B); B	12	0-11
Pitch of Tonic	C; $^\sharp$C (^{b}D); D; $^\sharp$D (^{b}E); E; F; $^\sharp$F (^{b}G); G; $^\sharp$G (^{b}A); A; $^\sharp$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

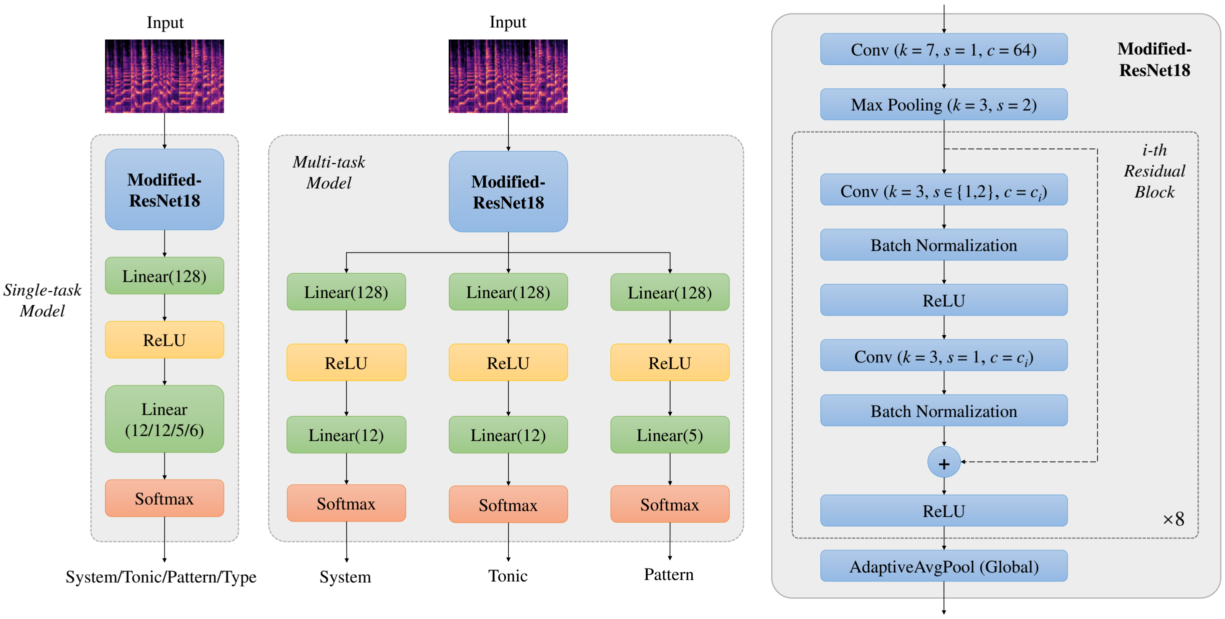

Figure 2. Model architectures. The shape of input is [Batch_size, Channel=1, F=168, T]. There are 4 single-task models for predicting TongGong system, tonic pitch, mode pattern and type respectively. Specifically, the single-task model for tonic prediction has no linear layer of 128 neurons and the following 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,

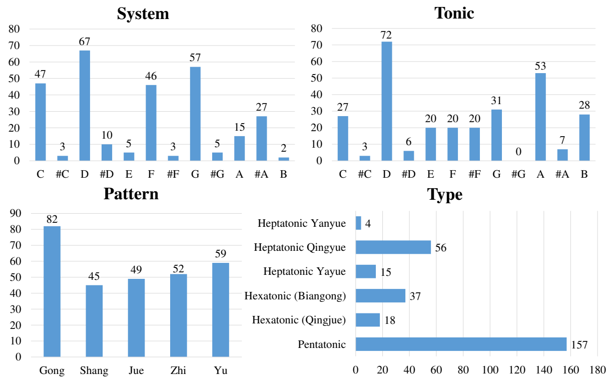

Figure 3. Data distribution of the CNPM Dataset

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.

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.

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

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.

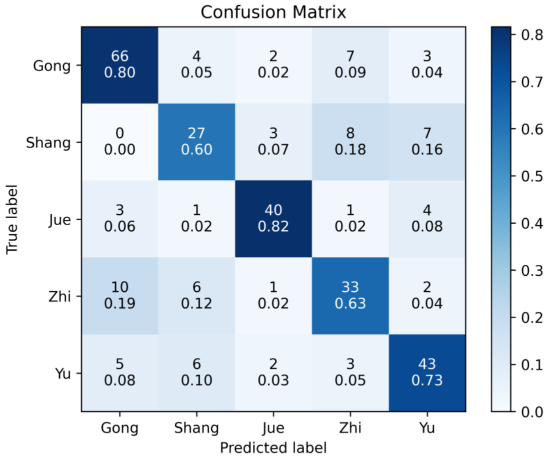

Figure 4. Normalized confusion matrix for mode pattern recognition results using the multi-task CNN model.

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.

TEACH YOURSELF GEORGIAN FOLK SONGS DATASET: AN ANNOTATED CORPUS OF TRADITIONAL VOCAL POLYPHONY

David Gillman
New College of Florida
dgillman@ncf.edu

Uday Goyat
Georgia Institute of Technology
ugoyat3@gatech.edu

Atalay Kutlay
New College of Florida
atalay.kutlay18@ncf.edu

ABSTRACT

New datasets of non-Western traditional music contribute to the development of knowledge in MIR and allow computational techniques to inform ethnomusicology. We present an annotated dataset of traditional vocal polyphony from two regions of the Republic of Georgia with disparate musical characteristics. The audio for each song consists of four polyphonic recordings of one performance from different microphones. We present a process and workflow that we use to annotate the dataset, which takes advantage of the salience of individual voices in each recording. The process results in an f_0 estimate for each vocal part.

1. INTRODUCTION

To evaluate algorithms in Music Information Retrieval (MIR) it is essential to have a variety of extensive datasets of annotated music [1]. Annotated datasets of vocal *a cappella* music have been scarce until recently, particularly of non-Western music. The work of [2] provides a list of datasets of vocal polyphony that includes two datasets of songs from the Republic of Georgia: the Erkomaishvili dataset of [3] and the collection recorded in the work of [4].

Multi-f_0 estimation is a sub-problem of Automated Music Transcription (AMT) consisting of identifying the fundamental frequency f_0 of each part in a polyphonic recording. Datasets of polyphony that are annotated with the fundamental frequency of each part are useful as ground truth for multi-f_0 algorithms. Multi-f_0 estimation is a challenging problem for *a cappella* vocal music because of the variety of sounds produced by the human voice and the similarity in timbre of different voices [5].

In this work we present an annotated dataset of 38 three-part songs from the Republic of Georgia, including 29 from the region of Guria and nine from the region of Samegrelo. The total duration of the collection is 89 minutes. Gurian songs are a particular challenge for multi-f_0 estimation. The three parts are independent melodic lines that often cross and contain rapid movement. The top part often consists of *krimanchuli*, Georgian yodeling, with as many as six changes of octave per second. These features make our dataset a useful contribution as part of a training set for multi-f_0 algorithms. We have created a web-based visualization of the dataset that is potentially useful as an aid to singers learning their parts on Georgian songs, which was the purpose of the original recordings.

We also present a new process and workflow for multi-f_0 estimation where several recordings exist of a single performance. Our dataset is an unusual challenge for annotation in that it does not include isolated tracks for each vocal part. Instead there are four recordings of one performance made from different microphones. One recording presents a balanced mix of voices. In each of the other three recordings one of the voices is more salient than the other two, but all three voices are easily audible and create a polyphonic mixture. Our process and workflow consist of isolating the salient voice, applying several algorithms for monophonic f_0 estimation, and using a graphical interface to select the correct estimate. In addition to f_0 estimates we present the median absolute deviation of the estimates, a measure of confidence in the estimates.

Other interactive methods have appeared for extracting melody from audio. The work of [6] introduced the Tony software for monophonic audio. It presents to the user several pitch estimates generated by an early stage of the pYin algorithm, from which the user can select ranges of time and frequency. It is designed for ease of use and has many features such as octave correction [7]. The system of [8] designed for the Erkomaishvili dataset presents to the user a spectrogram with one melody highlighted as a result of dynamic programming performed on a set of salient frequencies at each time step, followed by automatic corrections that take into account musical knowledge such as voice ranges. The user is able to delete and replace lines in the spectrogram. There is a web interface for the Erkomaishvili dataset which plays the audio of each song accompanied by a scrolling score with lyrics. The work of [2] created the Dagstuhl dataset by recording each singer with a larynx microphone, a headset microphone, and a dynamic microphone. The researchers applied both the pYin and CREPE algorithms to each recording and derived confidences for each algorithm on each microphone using a subset of recordings manually annotated by a sound engineer who used Tony.

Our method differs from these methods in that it is designed for polyphonic recordings that contain one salient voice, it makes use of several f_0 estimates, and it presents

Attribution: D. Gillman, U. Goyat, and A. Kutlay, "*Teach Yourself Georgian Folk Songs* Dataset: An Annotated Corpus Of Traditional Vocal Polyphony", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

a measure of confidence in the estimates. We have created a web interface for the dataset which plays the audio of each song accompanied by a scrolling visualization of the pitches. Unlike that of [2] our visualization shows the f_0 estimates of the three vocal parts, and the median absolute deviation of the each estimate is displayed in the manner of a confidence interval.

1.1 Georgian music datasets

The Republic of Georgia, bounded on the north by the Caucasus Mountains and on the west by the Black Sea, contains within its borders several starkly varying traditions of three-part *a cappella* vocal music [9]. Georgian singing is an aural tradition that uses a scale of its own [10]. The intervals that make up the Georgian scale are an area of research in ethnomusicology, as well as the extent to which singers in the tradition adjust to one another and deviate from a fixed scale to produce desired harmonies [11, 12]. Annotated datasets of Georgian music have been useful in advancing this research [9, 13].

The Erkomaishvili dataset contains Georgian sacred songs performed by a single performer on all three parts, with overdubbing. The most recent version of the dataset due to [3] includes f_0 annotations due to [8], onset annotations, and musical notation due to [14]. The dataset due to [4] includes Georgian folk songs from various regions recorded using individual larynx and headset microphones and a microphone for the ensemble. The authors observe that larynx microphones isolate the three parts well, and they illustrate this point by showing estimates of f_0 derived for one song in the collection.

The present dataset, like these other datasets, consists of recordings of performances by expert Georgian singers of interest to ethnomusicologists. However, unlike the dataset due to [3], our dataset consists of studio recordings of live ensembles, and unlike the dataset due to [4], our dataset was recorded without the benefit larynx microphones and therefore presents a greater challenge for pitch estimation.

Our dataset contains the recordings in the collections *Let Us Study Georgian Folk Songs (Gurian Songs)* and *Teach Yourself Georgian Folk Songs - Megrelian Songs*, both published in 2004 by the International Centre for Georgian Folk Song. The performers are highly trained musicians and experts in the music of their regions. [1] The performers were recorded together as an ensemble in close proximity in a single room in a studio [15]. There was one microphone for each part and one microphone for the room. This system of recordings was intended as an aid for singers learning their parts on Georgian songs, a setting in which it is beneficial to hear all three parts, but one above the other two. All songs are in three parts except during short intervals of overlap between antiphonal ensembles or solo and trio. The ensemble for each song is one high tenor (*pirveli* in Georgian), one middle tenor (*meore*) and one bass (*bani*), with a few exceptions. The song "Khasanbegura" has two antiphonal ensembles, one a trio and the other consisting of one high tenor and a choir of middle tenors and basses. The three songs "Maq'ruli", "Orira", and "Shvidk'atsa" also have two ensembles, one a trio and the other consisting of two tenors and a bass choir. Two songs, "Indi-Mindi" and "P'at'ara Saq'varelo", have a fourth part, a brief solo bass, that we have ignored for purposes of f_0 estimation. All but the two songs "Sabodisho" and "Mi Re Sotsodali" are *a cappella*. The accompaniment in these songs is a *chonguri*, a picked Georgian lute.

1.2 f_0 Estimation

Research on monophonic f_0 estimation has developed over several decades. The work of [16] introduced the use of the "cepstrum" in f_0 estimation, which exploits the fact that harmonics are regularly spaced in the frequency domain. Building on the work of [17], which exploited the same fact using the technique of spectral compression, the work of [18] introduced the technique of subharmonic summation to model the ability of the human ear to detect a weak fundamental pitch and used numerical methods to reduce the susceptibility of the estimate to noise [19]. The work of [20] introduced an algorithm for f_0 estimation and voiced-unvoiced classification which used the autocorrelation function of the signal to generate candidate pitches and dynamic programming to generate a sequence of pitches with few large jumps in frequency and few isolated voiced or unvoiced points. The well-known Praat software uses this algorithm [21]. The work of [22] introduced the well-known Yin algorithm, which used the sum of squared differences between the signal and lagged signals instead of the autocorrelation function. The work of [23] introduced the use of the fast-lifting wavelet transform for f_0 estimation. The work of [24] introduced the use of a neural network for f_0 estimation with the CREPE algorithm.

We use these six algorithms – Noll [16], Hermes [18], Boersma [20], Yin, Maddox [23], and CREPE – as a "panel of experts" to estimate f_0 in these recordings. [2] One may anticipate that algorithms based on autocorrelation or deep learning are strictly better than those based on Fourier or wavelet analysis. However, we find that for the high-quality non-monophonic recordings of this dataset, each algorithm produces correct estimates that tend to persist through the duration of a musical note or longer, and the algorithms err under different conditions. The errors we observe are mainly pitches in the wrong octave and pitches on the wrong part, including pitch estimates at times when the current part is silent or is a fricative consonant. Despite the presence of such errors, on harmonic content (judged subjectively as vowels sung by a healthy voice) at least one of the algorithms typically estimates f_0 correctly.

The paper is arranged as follows. In Section 2 we give details about the data and describe the process and workflow. [3] In Section 3 we present visualizations that com-

[1] The performers on the Gurian songs include Guri Sikharulidze, Tristan Sikharulidze, Otar Berdzenishvili, Anzor Erkomaishvili, Gedevan Mzhavanadze, Kote Papava, and Levan Goliadze. The performers on the Megrelian songs include members of the Odoia Choir directed by Polikarpe Khubulava.

[2] pYin did not improve over Yin in initial testing.

[3] The code resides in a public Github respository [25] The raw data and annotations are available to researchers upon request.