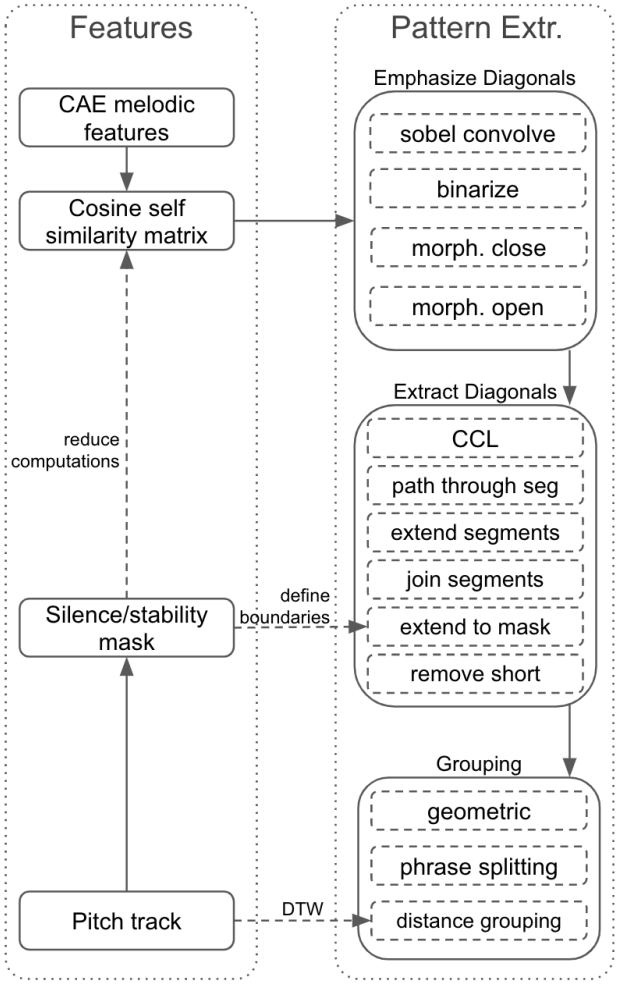

Figure 1. Pipeline overview - two feature sets are extracted from audio (pitch track and CAE features) and used to inform the extraction process.

3.1.1 Pitch Track and Mask

Recent work on vocal pitch extraction for Carnatic Music shows that state-of-the-art heuristic and data-driven methods for such tasks struggle to discriminate between singing voice and violin, and are not able to fully capture the vocal ornamentations in this tradition [29] **[CHAR7]**. To overcome this problem, a frequency-temporal attention network denoted FTA-Net [11] is re-trained using the SCMS dataset, further illustrating the positive impact this has for downstream tasks such as melodic pattern recognition. In this work we take advantage of this procedure to extract cleaner and more versatile pitch tracks from the audio in the Saraga dataset. We prioritize the close-microphone singing voice track for the extraction if available, however, there is no significant performance drop when running the Carnatic-tuned FTA-Net on top of mixed recordings. To account for incorrectly annotated silence that occur within ornaments due, for example, to glottal closure or other rapid vocal movements, we interpolate the pitch tracks to fill gaps shorter than or equal to 250ms [32].

As noted in [29] and [27], sañcāras are unlikely to contain long regions of sustained silence or pitch stability. We identify these regions in our pitch tracks and use this information to inform our search [**CHAR1**].

Silence is annotated as the regions of the pitch track that correspond to 0 Hz, in theory these are all regions where there is no sung voice (remembering that all silences of 250ms or less have been interpolated).

Stability is computed using the method outlined in [27, 29] - for a hop size of 0.2s, a window is passed across the entirety of the pitch track, those windows in which the maximum/minimum frequency is within $\pm$8Hz of it's mean are considered stable. Regions of the pitch track that contain consecutive stable windows whose total length sum to more than 1s are annotated as corresponding to a held note. The remaining windows, stable or not, are left unannotated. The final silence/stability mask, M_{ss}, annotates regions that correspond to silence or a held note with 1, and 0 otherwise.

3.1.2 Melodic Features and Self-similarity

We extract melodic features using a Complex Autoencoder (CAE). Mapping a signal, x, onto complex basis functions learnt by the CAE results in a transformation-invariant "magnitude space", $\mathbf{r}_x$, and a transformation-variant "phase space", ϕ_x. Exploiting the invariance-property of $\mathbf{r}_x$ has proven to achieve state-of-the-art results in repeated section discovery for audio [10].

We train a CAE on the vocal tracks in SCV, represented by their constant-Q transformed spectrogram with a hop size of 1984. The range comprises 120 frequency bins (24 per octave), starting from a minimal frequency of 80 Hz (selected as the minimum frequency expected in Carnatic vocal performance [33]). The spectrogram is split into n-grams of 32 frames. Before training, samples from the training data are randomly transposed by [-24 .. 24] frequency bins or time shifts of [-12 .. 12].

Each audio recording in SCV is mapped onto the complex basis functions learnt by the CAE and the transformation invariant "magnitude space", $\mathbf{r}_x$, used as melodic features relevant to the task of repeated pattern discovery.

As in [10], we compute a self-similarity matrix, $\mathbf{X}$, of $\mathbf{r}_x$, using the reciprocal of the cosine distance (plus some negligible epsilon to avoid division by zero). With some Carnatic performances lasting upwards of one hour, this computation rapidly becomes expensive. We use the masks learnt in Section 3.1.1, M_{ss}, to reduce the computation to include only those areas corresponding to regions we are interested in, reducing the number of pairwise comparisons by around 67%. Typically, $\mathbf{X}$ is between around 10,000 and 50,000 elements squared **[CHAR1, CHAR6]**.

Diagonals in $\mathbf{X}$ of high value correspond to two regions of continued similarity (repeated pattern). It is these diagonals we want identify and extract.

3.2 Repeated Pattern Extraction

Strong diagonal segments in $\mathbf{X}$ correspond to two occurrences of a repeated pattern. We define and extract them using a series of image processing techniques.

3.2.1 *Emphasizing diagonals*

To accentuate the diagonal segments, $\mathbf{X}$ is first convolved with a Sobel Filter to emphasize edges and then binarized about some experimentally set threshold, T_B. Since the emphasized edges correspond to the borders of a diagonal segment and not the segment itself, the binary matrix is morphologically closed to fill gaps and morphologically opened to smooth and remove noisy artifacts. Figure 2 illustrates this process.

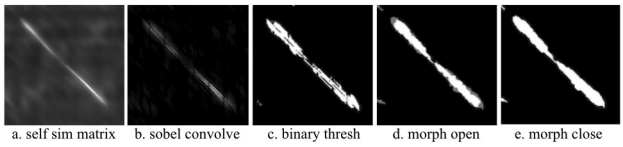

Figure 2. Emphasizing diagonals with convolution and morphological transformations: (a) Self-similarity matrix, (b) convolve with sobel filter, (c) binarize, (d) morphological opening, (e) morphological closing.

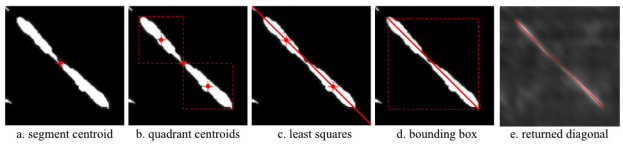

Figure 3. Defining diagonals from non-zero regions. Each segment is a thick region of non-zero values in $\mathbf{X}$, we extract a single line through this region corresponding to the underlying diagonal segment: (a) Identify segment centroid, (b) identify quadrant centroids, (c) least squares line, (d) identify bounding box, (e) return diagonal within bounding box.

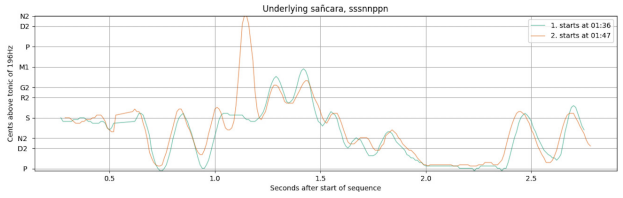

Figure 4. Two instances of the underlying sañcāra, sssnnppn, corresponding to the diagonal segment in Figures 2 and 3

3.2.2 *Extracting Diagonals*

Since instances of the same sañcāra can be performed at differing tempos, segments in $\mathbf{X}$ are not necessarily parallel to the $x = y$ diagonal, as such the method used to extract them in [10, 34] is not appropriate. We instead use a two-pass binary connected-components labelling algorithm (CCL) with a structuring element that considers elements that touch diagonally [35] to identify and group connecting non-zero elements, this can be conceptualized as a *flood fill* algorithm applied to all non-zero regions of $\mathbf{X}$. Each non-zero group corresponds to multiple x, y coordinates, we want to define a single path through this group to nominate as a candidate for the underlying true diagonal segment. First we compute the centroid of these coordinates and split the bounding box into an upper left quadrant and lower right quadrant centered on it. A line is learnt using least squares regression on the points connecting the two quadrant centroids and the underlying diagonal segment defined along that line between the bounding box of the group as a whole (Figure 3) **[CHAR5]**.

Segments are extended along their trajectory for a maximum of up to 50% of their length either side if the similarity in $\mathbf{X}$ for these regions is below some reduced threshold, T_{ext} where $T_{ext} < T_B$ (Figure 5) **[CHAR3]**.

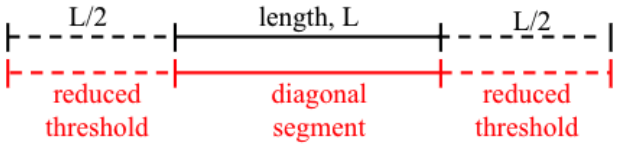

Figure 5. Reducing threshold for significant similarity in regions bordering returned segments. The two regions bordering the trajectory of a returned segment are subject to a lower threshold for similarity to account for variations in preceding and succeeding svaras in sañcāras.

Neighboring segments are joined if they are within 0.5s of each other and their gradients differ by $\leq$0.07rad **[CHAR4]**.

Segments that traverse regions identified in M_{ss} are broken along the center points of these silent/stable regions. Segments that are within 50% of their length to a silent or stable region (annotated in M_{ss}) are extended to the centroid of that silent/stable region **[CHAR1]**.

3.3 Grouping

Relationships between identified patterns are implicit in the geometry of $\mathbf{X}$ - each segment corresponds to two separate instances of the same pattern, those segments that intersect in x correspond to the same region in the recording (that represented on the x-axis) and those in y, vice versa. This information is used to iteratively group all patterns.

3.3.1 *Phrase Splitting*

We can further exploit segment positions in $\mathbf{X}$ to learn boundaries between shorter segments and longer phrases. If two segments intersect on one axis but are of different lengths, new segments are created corresponding to the intersecting portion of the segments **[CHAR2]**. Figure 6 illustrates an example of such. Segment A (corresponding to two patterns, $[x_1 : x_2]$ and $[y_1 : y_2]$) intersects with Segment B (corresponding to two patterns, $[x_1 : x_3]$ and $[y_3 : y_5]$). Consequently, two new segments are created from Segment B, $[y_3 : y_4]$ and $[y_4 : y_5]$ where $[y_3 : y_4]$ is another occurrence of $[x_1 : x_2]$. All patterns corresponding to Segment A, Segment B and the two new sub-segments are included in the final results.

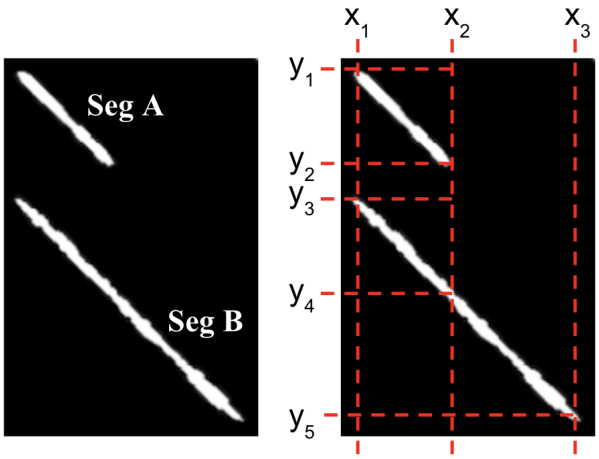

Figure 6. Learning sub-phrase relationships using geometry in $\mathbf{X}$ - Segment B is split into two since a subsequence of Segment B exists in isolation elsewhere (Segment A), both the entire Segment B and it's two subsequences are returned.

Performance	N_r	N_a	Recall	Precision	F1
KJ	174	168	0.73	0.76	0.74
SJ	214	106	0.53	0.54	0.52
VNK	90	144	0.33	0.42	0.37
Overall	478	418	0.54	0.60	0.57

Table 1. Melodic pattern discovery results for the three performances in Table 2. N_r and N_a are the number of returned and number of annotated patterns respectively

3.3.2 Distance Grouping

Further grouping is attempted by comparing each pairwise combination of groups. In each comparison, the dynamic time-warping distance is computed between 10 sets of two randomly selected patterns (one from each group) using the pitch tracks. If the average DTW distance between two groups is below some threshold, T_{dtw}, the groups are considered to represent the same pattern and merged to form one **[CHAR4]**.

4. EXPERIMENTAL SETUP

We apply our approach to three Carnatic performances (Table 2) and compare the results against expert annotations.

4.1 Annotations

Ground-truth annotations of all *sañcāras* in the audio recordings were created by a professional Carnatic vocalist who has 21 years of performance experience in South India. Annotations were created in Carnatic notation, known as *sargam* [36], using the software ELAN [37]. As there is no definitive list of *sañcāras* in a given *rāga*, the segmentations are based on the annotator's experience as a professional performer and student of a highly esteemed musical lineage. These annotations are therefore subjective to some degree, but have the benefit of being based on expert performer knowledge rather than on an externally imposed metric that may be irrelevant to musical concepts held by culture bearers.

In *kritis*, one of the most popular compositional formats, initial phrases are repeated with variations, known as *saṅgatis*. This means that there will be initial *sañcāras* that are subsequently varied. These musical connections are captured by grouping the related musical material together with an 'underlying *sañcāra*' annotation, which refers to the first occurrence and typically simplest version of the *sañcāra*. To reflect the hierarchical nature of plausible musical segmentations, longer full phrase-level annotations that can comprise several shorter *sañcāras* are also created **[CHAR1]**. The evaluation reported here is made on the 'underlying *sañcāra*' and 'underlying full phrase' levels.

4.2 Method and metrics

The various parameters at every step are optimized on one performance, Koti Janmani, representing 35% of the total number of minutes performed across the three performance. Those parameters related to the diagonal segment extraction are selected so as to maximize recall in identifying 'underlying *sañcāra*s' and 'underlying full phrases' that occur more than once in the ground-truth annotations. To consider whether a returned pattern, *R*, is a match with an annotation, *A*, we consider the overlap between them. Since the ideal boundary between shorter segments within a longer phrase is not always clear cut **[CHAR2]**, the expectation of 100% overlap is unrealistic **[CHAR1]**. As in [29], we consider *A* and *R* to be a match if the intersection between them is more than two-thirds the length of *A and* more than two-thirds the length of *R*, inspection of results find this limit to be sufficient in identifying and exploring regions of melodic interest.

5. RESULTS

Table 1 presents the results for the three performances. Our process is able to find 226 of the 418 annotations across the three recordings, corresponding to a total recall of 0.54, precision of 0.60 and F_1 of 0.57. The reader is encouraged to browse and listen to the results using the online visualisation tool provided in our Github repository.

6. DISCUSSION

Although far from 100 percent of annotated patterns were found, we significantly exceed what would have been achieved if we had used out-of-the-box methods without any consideration for the specifities of the Carnatic traditon. For example, when applying the pattern extraction process presented in [10] using our CAE model trained on the SCV, we are able to identify only 8% of annotated patterns at a precision of 7%. It is unsurprising that our performance exceeds this figure since our process is built, in part, on the same underlying model, replacing the pattern extraction steps with more tradition-informed ones.

The results achieved could contribute to a useful tool for musicologists needing to find many instances of the

Alias	Artist	Title	*Rāga*	Composer	Duration
KJ	Akkarai Sisters	Koti Janmani	*Ritigowla*	Oottukkadu Venkata Kavi	08:46
SJ	Salem Gayatri Venkatesan	Sharanu Janakana	*Bilahari*	Purandara Dasa	07:02
VNK	Sumitra Vasudev	Vanajaksha Ninne Kori	*Ritigowla*	Veenai Kuppaiyer	08:37
				Total:	24:25

Table 2. Three Carnatic performances used for evaluation.

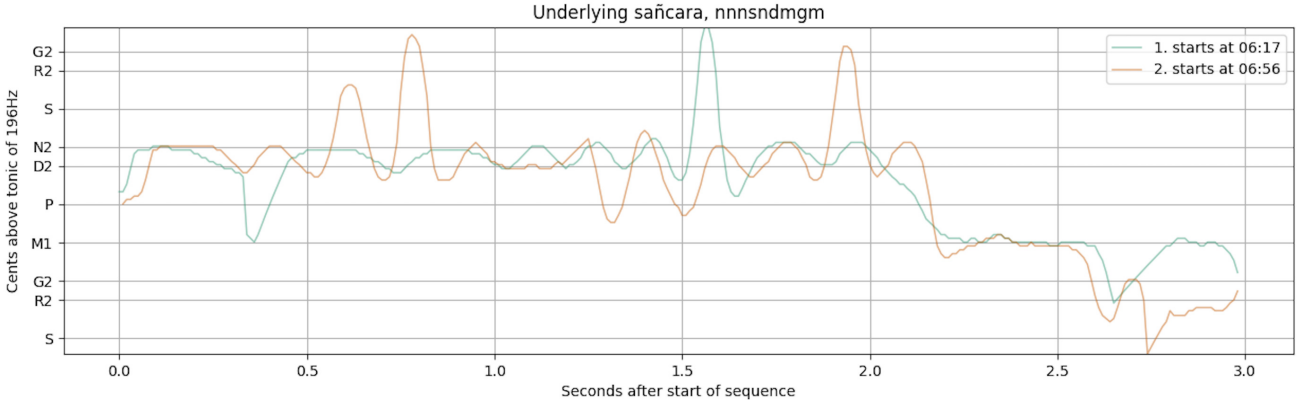

Figure 7. The pattern "nnnsndmgm" in Koti Janmani. The underlying *sañcāra* (blue line) is performed "nnnsndmgm", then later as a distinct variation "nnssndpdnndmmgrgm" (orange line).

same *sañcāra* across a long or many performances, a process that would otherwise involve extremely time consuming labour. While yet to be evaluated formally, the nature of the groupings returned appear promising from a musicological perspective. Figure 4 displays the pitch tracks for the two patterns corresponding to the diagonal segment in Figures 2 and 3. Each pattern would be considered instances of the same underlying sañcāra, "sssnnppn", but in fact are realized slightly differently ("snsgrsnnpn" at 01:36 and "snsmgrsnnpn" at 01:47). The steps outlined in Figures 2 and 3 illustrate how this subtle variation in performance is captured by considering the structural features that are particular to this musical style. However, there comes a point where the variation performed on the underlying or initial *sañcāra* is too great to be found by the process, and indeed it would also not be considered the same *sañcāra* by musicians, but rather a variation or new *saṅgati*. For example, in Koti Janmani, the initial *sañcāra* "nnnsndmgm" is later performed as a distinct variation "nnssndpdnndmmgrgm", and these two are successfully separated by the matching process into two motif groups (11 and 12) (Figure. 7).

While in the *kritis*, Koti Janmani and Sharanu Janakana, the majority of patterns annotated are found, fewer were found in the *varṇam*, Vanajaksha Ninne Kori. This might be explained by the differing structural features of the two compositional styles and of these particular compositions. This *varṇam* has a long 28 beat metrical cycle that lasts approximately 22 seconds. Furthermore, the melodic line has a more discursive and wandering character than the two *kritis* in this analysis. *Kritis* are based on a theme and variation (*saṅgati*) structure, while *varṇams* typically have fewer repetitions of any given phrase. As a result of these qualities, it might be harder for our process to find the borders of segments that match with those of the annotator. As discussed, there is often more than one plausible option for segmentation within longer units. It is possible that even expert annotators would find a larger number of plausible segmentation options in this *varṇam*, something that we could test for in future work by asking two or more annotators to annotate all plausible options, rather than only those that seem optimal **[CHAR2]**.

To demonstrate how our results could be useful to musicologists and fans of the style, we provide a high-level tool to explore and listen to them. We also include the results for 6 more performances across 6 ragas not mentioned here for a total of 9 performances across 8 distinct ragas. We do not have expert annotations for these extra performances and hence cannot empirically evaluate the results. One important factor in how valuable such a tool is, is the extent to which retrieved *sañcāras* are effectively sorted into groups of high melodic similarity that matches musicians' concepts of musical similarity in the style. It is obvious when exploring the results, that even for a listener with little musical background, there is a strong degree of similarity between patterns returned in the same group, both for those patterns that correspond to annotations and otherwise.

7. CONCLUSION

By considering and incorporating the characteristic features of a musical tradition into the development process, we obtain musicologically relevant results for the task of repeated melodic pattern recognition in Carnatic Music. In an effort to provide value for researchers, musicologists and fans of the Carnatic Music tradition, we provide all code, data and a high-level visualisation tool to explore the results.

8. ACKNOWLEDGEMENTS

Thanks to the Carnatic vocalist, Brindha Manickavasakan for her considered annotations of the audio recordings.

This research work was carried out under the project Musical AI - PID2019- 111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

9. REFERENCES

[1] X. Serra, "A multicultural approach in music information research," *In Proc. of the 14th Int. Society for Music Information Retrieval Conf., (ISMIR),* Miami, Florida, pp. 151–156, 2011.

[2] T. Viswanathan, "The analysis of rāga ālāpana in South Indian music," *Asian Music*, vol. 9, no. 1, pp. 13–71, 1977.

[3] T. M. Krishna and V. Ishwar, "*Carnatic music: Svara, gamaka, motif and raga identity*," *2nd CompMusic Workshop*, pp. 12–18, 2012.

[4] L. Pearson, "Coarticulation and gesture: an analysis of melodic movement in South Indian raga performance," *Music Analysis*, vol. 35, no. 3, pp. 280–313, 2016.

[5] V. S. Viraraghavan, A. Pal, H. Murthy, and R. Aravind, "State-based transcription of components of carnatic music," *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 811–815, 2020.

[6] H. G. Ranjani, D. Paramashivan, and T. V. Sreenivas, "Quantized melodic contours in indian art music perception: Application to transcription," *In Proc. of the 18th Int. Society for Music Information Retrieval (ISMIR),* Paris, France, 2017.

[7] H. G. Ranjani, D. Paramashivan, and T. Sreenivas, "Discovering structural similarities among rāgas in indian art music: a computational approach," *Sādhanā*, vol. 44, 05 2019.

[8] L. Pesch, *The Oxford illustrated companion to South Indian classical music*. Oxford University Press, USA, 2009.

[9] V. Ishwar, S. Dutta, A. Bellur, and H. A. Murthy, "*Motif Spotting in an Alapana in Carnatic Music.*" *In Proc. of the 14th Int. Society for Music Information Retrieval Conf. (ISMIR),* Curitiba, Brazil, pp. 499–504, 2013.

[10] S. Lattner, A. Arzt, and M. Dörfler, "Learning complex basis functions for invariant representations of audio," *In Proc. of the 20th Int. Society for Music Information Retrieval Conf. (ISMIR),* Delft, The Netherlands, 2019.

[11] S. Yu, X. Sun, Y. Yu, and W. Li, "Frequency-Temporal Attention Network for Singing Melody Extraction," *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Toronto, Canada, pp. 251–255, 2021.

[12] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for Music Information Research in Indian Art Music," *Int. Computer Music Conf./Sound and Music Computing Conf.,* Athens, Greece, pp. 1029–1036, 2014.

[13] D. Deutsch, "Grouping mechanisms in music," in *The psychology of music*. Academic Press, 1982, pp. 99–134.

[14] J. Tenney and L. Polansky, "Temporal gestalt perception in music," *Journal of Music Theory*, vol. 24, no. 2, pp. 205–241, 1980.

[15] C. Hasty, "Segmentation and process in post-tonal music," *Music Theory Spectrum*, vol. 3, pp. 54–73, 1981.

[16] A. Volk and P. van Kranenburg, "Melodic similarity among folk songs: An annotation study on similarity based categorization in music," *Musicae Scientiae*, vol. 16, no. 3, pp. 317–339, 2012.

[17] P. Boot, A. Volk, and W. Bas de Haas, "Evaluating the Role of Repeated Patterns in Folk Song Classification and Compression," *Journal of New Music Research*, vol. 45, no. 3, pp. 223–238, 2016.

[18] P. Rao, J. C. Ross, and K. K. Ganguli, "Distinguishing raga-specific intonation of phrases with audio analysis," *Ninaad*, vol. 26, no. 1, pp. 59–68, 2013.

[19] I. Y. Ren, A. Volk, W. Swierstra, and R. C. Veltkamp, "In search of the consensus among musical pattern discovery algorithms," *In Proc. of the 18th Int. Society for Music Information Retrieval (ISMIR),* Paris, France, pp. 671–680, 2017.

[20] B. Janssen, W. Bas de Haas, A. Volk, and P. van Kranenburg, "Finding repeated patterns in music: State of knowledge, challenges, perspectives," *Lecture Notes in Computer Science*, no. 8905, pp. 277–297, 2014.

[21] M. Thomas, Y. Murthy, and S. G. Koolagudi, "Detection of largest possible repeated patterns in Indian audio songs using spectral features," *In Proc. of the IEEE Canadian Conf. on Electrical and Computer Engineering (CCECE)*, 2016.

[22] C. Wang, J. Hsu, and S. Dubnov, "Music Pattern Discovery with Variable Markov Oracle: A Unified Approach to Symbolic and Audio Representations," *In Proc. of the 16th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 176–182, 2015.

[23] G. Sankalp, J. Serrà, V. Ishwar, and X. Serra, "Mining melodic patterns in large audio collections of indian art music," *In Proc. of the Int. Conf. on Signal Image Technology and Internet Based Systems (SITIS),* Marrakech, Morocco, pp. 264–271, 2014.

[24] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. A. Murthy, "Classification of Melodic Motifs in Raga Music with Time-series

Matching," *Journal of New Music Research*, vol. 43, no. 1, pp. 115–131, 2014.

[25] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. Murthy, "Melodic motivic analysis of indian music," *Journal of New Music Research*, vol. 43, no. 1, pp. 115–131, 2014.

[26] S. Dutta and H. A. Murthy, "Discovering typical motifs of a raga from one-liners of songs in Carnatic Music," *In Proc. of the 15th Int. Society for Music Information Retrieval*, pp. 397–402, 2014.

[27] T. Nuttall, G. Plaja-Roglans, L. Pearson, and X. Serra, "The matrix profile for motif discovery in audio - an example application in carnatic music," *In Proc. of the 15th Int. Symposium on Computer Music Multidisciplinary Research (CMMR),* Tokyo, Japan, pp. 109–118, 2022.

[28] G. Padmasundari and H. A. Murthy, "Raga identification using locality sensitive hashing," *In Proc. of the 23th National Conf. on Communications (NCC)*, 2017.

[29] G. Plaja-Roglans, T. Nuttall, L. Pearson, and X. Serra, "Repertoire-specific vocal pitch data generation for improved melodic analysis of carnatic music," 2022. [Online]. Available: 10.5281/zenodo.7036117

[30] A. Lele, S. Pinjani, K. K. Ganguli, and P. Rao, "Improved melodic sequence matching for query based searching in Indian Classical Music," *In Proc. of the Frontiers in Signal and Music Processing (FRSM)*, 2016.

[31] A. Srinivasamurthy, S. Gulati, R. Repetto, and X. Serra, "Saraga: Open datasets for research on indian art music," *Empirical Musicology Review*, 2020.

[32] S. Gulati, J. Serrà, K. Ganguli, S. Şenturk, and X. Serra, "Time-delayed melody surfaces for Raga recognition," *In Proc. of the 17th Int. Society for Music Information Retrieval Conf. (ISMIR),* New York City, USA, pp. 751–757, 2016.

[33] M. Venkataraman, P. Boominathan, and A. Nallamuthu, "Frequency Range Measures in Carnatic Singers," *Journal of Voice*, 2020.

[34] S. Lattner, M. Grachten, and G. Widmer, "Learning transposition-invariant interval features from symbolic music and audio," *In Proc. of the 19th Int. Society for Music Information Retrieval Conf. (ISMIR),* Paris, France, 2018.

[35] J. R. Weaver, "Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors," *The American Mathematical Monthly*, vol. 92, no. 10, pp. 711–717, 1985.

[36] N. Ramanathan, "Sargam and musical conception in karnataka system'," *Sargam as a musical material*, 2004.

[37] H. Sloetjes and P. Wittenburg, "*Annotation by category-ELAN and ISO DCR*," Paper presented at the6th Int. Conf. on Language Resources and Evaluation (LREC), May, 2008.

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.