# Raga Detection – A Comparison Between Transfer Learning and Audio Processing

Varsha Venkatakrishnan[1], Abirami Vellingiri Thirunavukkarasu[2], Sampath Kumar P[3]

[1] Department of Computer Science and Engineering, PSG College Of Technology, Coimbatore, India
[2] Department of Computer Science and Engineering, PSG College Of Technology, Coimbatore, India
[3] Department of Computer Science and Engineering, PSG College Of Technology, Coimbatore, India

[1]varshakvenkat@gmail.com; [2]abiramigiri1998@gmail.com; [3]psk.cse@psgtech.ac.in

***Abstract*— Raga is one of the core concepts in Carnatic music. Determination of Raga is the critical concept in information retrieval concerning Carnatic songs. Each Raga plays a significant role in bringing out different emotions in the listener. When there is a need to create a particular atmosphere, it could be advantageous to choose Carnatic songs of a specific Raga. Consequently, Raga classification has significant utility. This paper summarises potential approaches to analyse the audio sample and predict the Melakarta Raga. Transfer learning and audio processing are the two different approaches discussed to detect Melakarta Raga. In transfer learning, knowledge is transferred from a variety of base tasks - instrument detection, spoken digit classification and genre classification to the Melakarta Raga detection model, thus exploring different ways to choose the optimal base task for this transfer learning problem. The audio processing approach uses individual note extraction and vocal range detection to obtain the set of Swaras used in the audio sample. Classification of Melakarta involves comparing the list of Swaras in the audio sample to the dataset comprising the combination of Swaras for each Melakarta.***

***Keywords*— Machine Learning, Transfer Learning, Raga Detection, Audio Processing, Pattern Detection***

## I. INTRODUCTION

Carnatic music is a traditional Indian art form which adheres to strict frameworks such as Raga and Thalam. The Raga of a song determines the pattern of Swaras it can use. The Melakarta Raga is the broadest classification of Raga, and there are 72 of them. A Melakarta consists of a unique combination of 8 Swaras out of the 12-note system. Every Carnatic song falling under a Melakarta contains the Swaras in that Melakarta, and so Melakarta detection is a classic pattern detection problem.

Audio processing and transfer learning are two ways considered in this paper to implement pattern detection and identify the Melakarta Raga of an audio sample. In transfer learning, the choice of base task has a crucial impact on the model's performance. The three chosen base tasks share similarities with the Melakarta detection task in different ways. For instance, the instruments making up the background music also adhere to the Raga like vocals in the Melakarta detection dataset. As a result, the instrument classification task is a good base task to train the model to learn from the instrumental track in the audio sample. The similarity between the spoken digit classification dataset and the Melakarta detection dataset lies in the fact that they both have vocal samples. Both genre classification in western music and Melakarta detection in Carnatic music are similar problems as they involve pattern detection. The following sections of the paper give specific details on the performance and methodologies of these approaches.

## II. AUDIO PROCESSING

Audio processing for Melakarta Raga detection involves extracting the frequencies from an audio sample and detecting patterns based on the rules of Raga to identify the Melakarta Raga of a song. Pre-processing the audio samples is followed by analysing the patterns of Swaras and extraction of vocals based on the REPET-SIM method of Rafii and Pardo, 2012 [5]. Swaras are the individual notes of music like "sa", "ri", "ga", "ma" which have different and incremental frequencies. A frame is the length of a segment of the audio file taken for processing. The vocal track of a song consists of similar frequencies in the frame in comparison to the background music, and it is separated from the rest of the audio sample using a filter. Fig. 1 shows the spectral distribution of the audio file before and after vocal extraction.

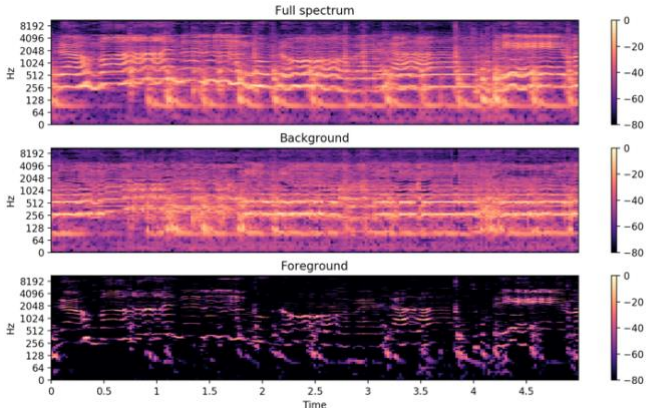


Fig. 1 Spectral distribution before and after vocal extraction

In the case of audio samples with no background music, there is no need for the extraction step. Significant portions of Carnatic songs lie within 13 consecutive frequencies, and there are several overlapping windows of these 13 frequencies. Within a window, every Melakarta has the Swaras “sa”, “pa” and “sa”, where each “sa” represents the extreme ends of frequencies in the song, and “pa” is the centremost frequency. Apart from the three, there are five more Swaras chosen for every Melakarta. There exist three frequencies each for the Swaras “ri”, “ga”, “dha” and “ni” and two frequencies for the Swara “ma”. Every Melakarta Raga has a unique combination of frequencies chosen to represent these 5 Swaras. Fig. 2 shows the keyboard layout for two different Melakarta.

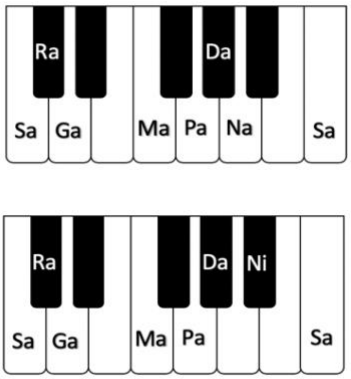


Fig. 2 Keyboard layout for Kanakangi and Ratnangi

## *A. Determination Of Vocal Range*

Identifying the vocal or instrumental range is done by isolating the window of frequencies with the maximum occurrence in the audio sample. Traversing every window and counting the instances of each frequency gives the window with the highest count. Identifying the window also gives the Kattai of the audio sample. Fig. 3 shows the keyboard layout for two different Kattai. (1 and 1.5)

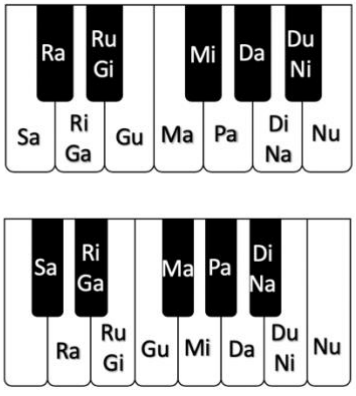


Fig. 3 Keyboard layout for 1 Kattai and 1.5 Kattai

*B. Determination Of Melakarta*

Identifying the vocal range is followed by isolation of the Swaras in the audio sample by comparing it to the Mel - frequency value of each key. The thirteen frequencies in the window are numbered from 1 to 13 and represent potential notes from lower “sa” to higher “sa”. From this, the eight frequencies denoting the eight Swaras are isolated based on the following algorithm.

- {1, 6, 13} is set as “sa”, “pa”, “sa” as they are present in every Melakarta
- From {2, 3, 4, 5} two with the highest count are chosen, and the lower frequency is set as “ri”, and the higher is set as “ga”.
- From {7, 8} the frequency with the highest count, is set as “ma”.
- From {9, 10, 11, 12} two with the highest count are chosen, and the lower frequency is set as “dha”, and the higher is set as “ni”.

Comparison of the combination of Swaras obtained from the algorithm to the Swara patterns of 72 Melakarta gives the Melakarta Raga of the audio sample.

*C. Dataset Description*

The samples used for testing the audio processing implementation of Raga detection consist of Carnatic songs belonging to Melakarta Ragas such as Shanmugapriya, Harikamboji, Hamsadhwani, Ratnangi, Kanakangi, and Dheerasankarabaranam. Some of the samples were recorded on the keyboard, some were recordings of live vocals and some were audio samples with both vocals, and background instrumental music by various artists. The accuracy is almost perfect in the case of a high-quality audio sample sung or played in perfect pitch. However, some songs may have vocal runs or Gamakkas, which may cause incorrect detection of frequencies due to the algorithm’s high sensitivity. In such cases, the algorithm classifies Raga incorrectly.

In Fig. 4, the audio samples indicated as “With BGM” are audio files of Carnatic songs with both vocal and the background music track. The foreground is extracted from these samples to get the “BGM removed” samples. In practical situations, a Raga detection system should be able to classify such files as well. The algorithm finds its limitations in these cases as an audio sample “With BGM” may have varying frequencies due to the background music track. Audio samples that are “BGM removed” have low-quality audio compared to the original file, and some remnants of the background track cause varying frequencies yet again. So a different approach is considered to tackle these use-cases. Pattern detection using machine learning may be a potential solution.

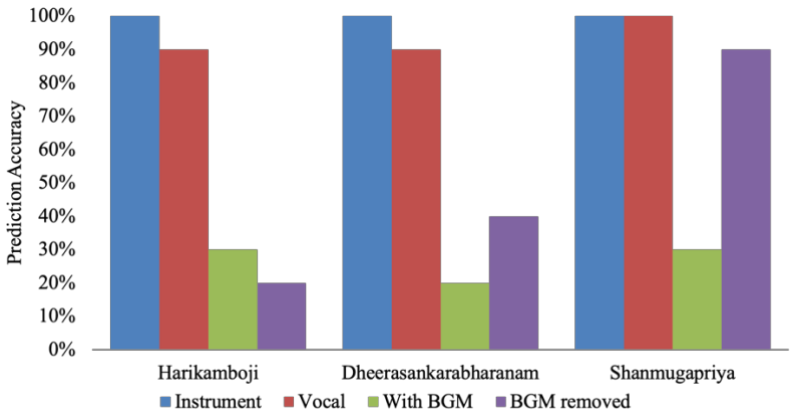


Fig. 4 Audio processing performance chart

## III. TRANSFER LEARNING

A traditional machine learning model used to solve the Raga Detection problem resulted in a very low accuracy for prediction. The observed poor accuracy could be related to the fact that Raga detection has no proper dataset that is easily accessible. Fig. 5 shows the accuracy of the machine learning model for raga detection.

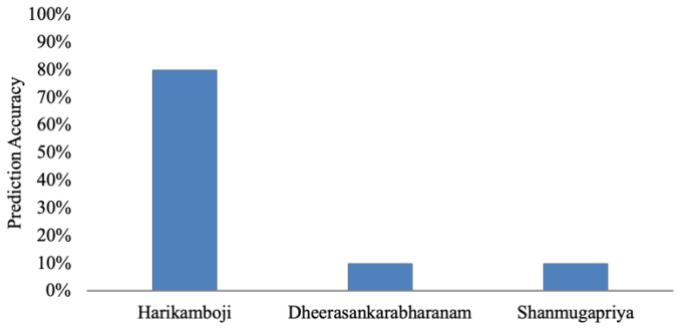


Fig. 5 Machine learning performance chart

Transfer learning is an experimental research technique in machine learning that facilitates solving problems that do not have varied and extensive datasets. Transfer learning operates by the transfer of knowledge from one problem to a different but relevant problem. The transfer of knowledge is through the distribution of weights in the edges between neurons. A suitable dataset is the foundation of a machine learning experiment, and Melakarta detection in Carnatic music does not have readily available labelled datasets, thus transfer learning is a practical approach. Even when there isn't a similar task to train as the base task, initial training with any dataset statistically outperforms random initialization of weights [7]. Construction of the base model for a data set of instrumental sounds, spoken digits and songs in western music belonging to different genres is followed by training the transfer model over the base model to detect Melakarta Raga in Carnatic music. MFCC is the choice of feature to extract information from the audio samples.

### *A. Dataset Description For Base Task*

Three different base tasks are considered, and the one that gives optimal results for transfer learning is determined. One base task is instrument classification using the Magenta Nsynth Dataset [4]. A collection of 4096 sound excerpts from the dataset is used for training. It has labelled data of instrumental sounds such as bass, keyboard, guitar, organ, string, brass, reed, mallet, flute and vocal notes. The distribution of different labels in the dataset can be found in Fig. 6.

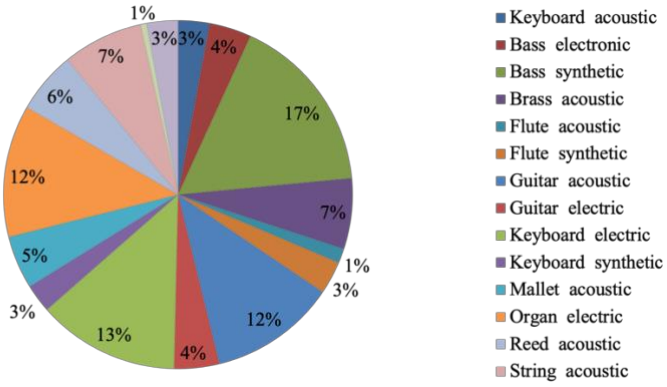


Fig. 6 Instrument Base task dataset distribution

The second base task is spoken digit classification using the Jakobovski/free-spoken-digit-dataset: v1.0.8 [6]. A collection of 3000 voice recording samples of digits from 0-9 spoken by three different speakers with varied accents, is used for training. The distribution of different labels in the dataset can be found in Fig. 7.

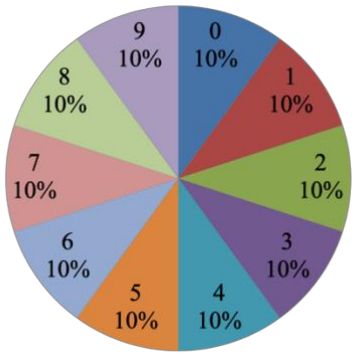


Fig. 7 Spoken digit Base task dataset distribution

The genre classification dataset for western music using the GTZAN dataset forms the building blocks of the third base task. A collection of 1000 samples of songs from 10 different genres that are among blues, classical, country, jazz, disco, hip-hop, metal, pop, reggae and rock. Fig. 8 illustrates the distribution of samples among different genres in the dataset.

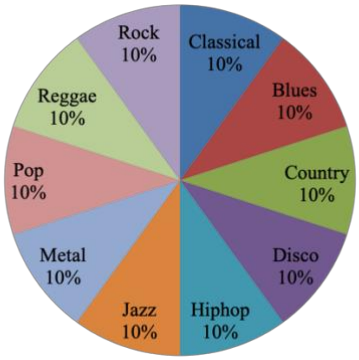


Fig. 8 Genre Base task dataset distribution

### *B. Model Description Of Base Task*

The model used for all three base tasks is a multilayer perceptron with a single hidden layer consisting of 45 neurons and an input layer with 40 neurons. The output layer of the instrument classification task, spoken digit classification task and genre classification task have 16, 10 and 10 neurons, respectively. The input layer consists of 40 neurons to accept 40 Mel Frequency Cepstrum Coefficient values (derived based on the frequency distribution of the audio sample) that are extracted for every audio sample. The output layer has 16 or 10 neurons depending on the base task to classify the input among its labels. Hyperparameters like the number of hidden layers, the number of neurons in the hidden layer, epochs and the learning rate using which the model is trained are determined based on their performance over a range of values.

### *C. Dataset Description For Transfer Task*

Audio samples were manually downloaded for Carnatic songs belonging to three Melakarta Raga (Dheerasankarabharanam, Harikamboji and Shanmugapriya). A total of 187 files of length 30 seconds each. The distribution of different labels in the dataset can be found in Fig. 9.

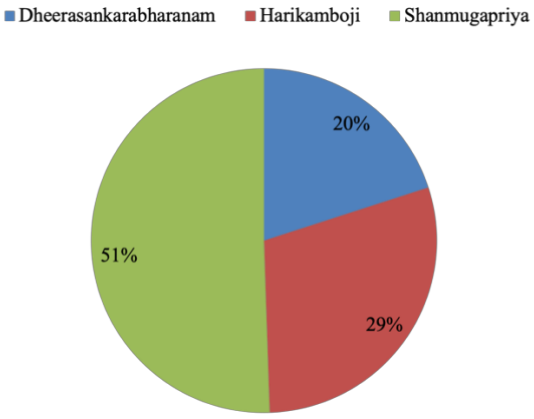


Fig. 9 Transfer task dataset distribution

### *D. Model Description Of Transfer Task*

The model from the base task is retained as such except for the output layer. Since the dataset consists of songs classified into 3 Melakarta Ragas, the output layer is modified to have 3 neurons instead of 10 or 16 from the base task. The Mel Cepstrum Frequency Coefficients are determined again for the transfer task's dataset and fed to train the model. Fig. 10 shows the performance metrics of transfer learning with each base task.

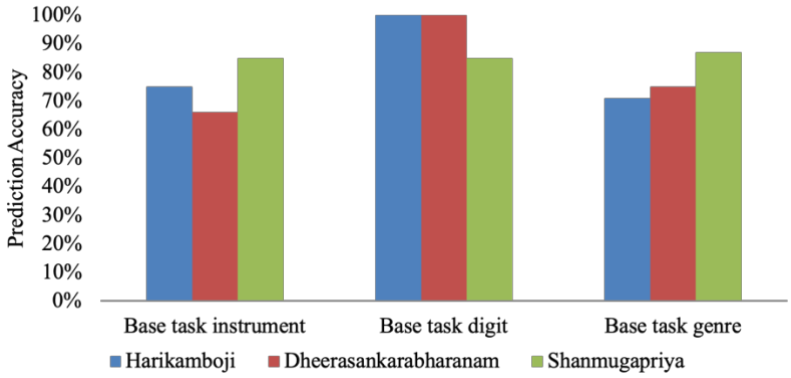


Fig. 10 Transfer learning performance chart

## IV. CONCLUSIONS

The two approaches used, though different in implementation, still follow the same concept of pattern detection. Audio processing is a practical approach for samples with stellar audio quality where it is possible to retrieve exact frequencies. In real-world situations, audio samples may not always have pristine quality, and so machine learning is an alternative approach for such cases. The lack of data gives poor prediction results in machine learning. It is overcome by opting for a machine learning technique - transfer learning where the lack of data is combated by initial training with a different yet similar dataset.

The spoken digit classification task outperformed the other two base tasks despite the larger size of the instrument classification dataset, indicating that size is not always the primary factor to be considered when choosing an optimal base task. The observed difference in performance could be because most of the information regarding Raga from the Carnatic music audio samples are from the vocal track and the spoken digit classification task is the one that shares the most similarities with the vocal track in comparison to the other two base tasks. It can be inferred from the results of this experiment that when choosing a base task, it can be advantageous to isolate the most important characteristic of the transfer task's dataset and locate a base task which is similar in that aspect.

Melakarta detection using audio processing was implemented for 72 Melakartas, and transfer learning was implemented only for 3 Melakartas. Audio processing outperforms transfer learning model described in this paper in terms of the number of Melakartas it can detect. However, transfer learning is more suitable for a broader range of samples and has fewer restrictions on the type of audio files it can classify. Moreover, in complex Ragas with less clarity in their structure, it is highly complex to detect the Raga of a song with audio processing algorithmically. In such cases, transfer learning is a better solution in comparison to audio processing. This research can be extended to detect such complex Raga by adding to the transfer learning model's design.

## REFERENCES


[1] Belle, Shreyas and Rushikesh Joshi. Raga Identification by using Swara Intonation. (2010).

[2] P.Sriram. Karnatic Music Primer. The Carnatic Music Association of North America, Inc. 1990.

[3] De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111(4), 1917-1930.

[4] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. 2017.

[5] Zafar Rafii and Bryan Pardo. Repeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. IEEE Transactions on Audio, Speech, and Language Processing, 21(1):71–82, January 2013.

[6] Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, & Adhish Thite. (2018, August 9). Jakobovski/free-spoken-digit-dataset: v1.0.8 (Version v1.0.8). Zenodo. http://doi.org/10.5281/zenodo.1342401

[7] Yosinski J, Clune J, Bengio Y, et al (2014) How transferable are features in deep neural networks?[J]. Eprint Arxiv 27:3320–3328