# SARAGA AUDIOVISUAL: A LARGE MULTIMODAL OPEN DATA COLLECTION FOR THE ANALYSIS OF CARNATIC MUSIC

**Adithi Shankar** **Genís Plaja-Roglans** **Thomas Nuttall**
**Martín Rocamora** **Xavier Serra**
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
adithishankar.sivasankar@upf.edu

## ABSTRACT

Carnatic music is a style of South Indian art music whose analysis using computational methods is an active area of research in Music Information Research (MIR). A core, open dataset for such analysis is the Saraga dataset, which includes multi-stem audio, expert annotations, and accompanying metadata. However, it has been noted that there are several limitations to the Saraga collections, and that additional relevant aspects of the tradition still need to be covered to facilitate musicologically important research lines. In this work, we present Saraga Audiovisual, a dataset that includes new and more diverse renditions of Carnatic vocal performances, totalling 42 concerts and more than 60 hours of music. A major contribution of this dataset is the inclusion of video recordings for all concerts, allowing for a wide range of multimodal analyses. We also provide high-quality human pose estimation data of the musicians extracted from the video footage, and perform benchmarking experiments for the different modalities to validate the utility of the novel collection. Saraga Audiovisual, along with access tools and results of our experiments, is made available for research purposes.

## 1. INTRODUCTION

In recent years, there has been an increasing emphasis on representing non-Western classical music styles within computational musicology [1, 2], an interdisciplinary research area involving musicology and computer science. To facilitate this research, many repertoire-specific datasets have been proposed that take into account the melodic, rhythmic and structural complexities of these traditions. Several of them are consolidated within the scope of Dunya, a collection of large music corpora dedicated to fuelling research of five major non-Western music traditions: Carnatic music, Hindustani music, Turkish Makam, Beijing Opera and Arab-Andalusian music [1].



One style of particular interest is Carnatic music, for which there has been numerous computational musicological studies carried out using the Saraga dataset, a subset of the Dunya corpora dedicated to the Indian Art Music (IAM) traditions of Hindustani and Carnatic music [3–6]. The Carnatic portion of this dataset comprises performance audio, expert/automatically extracted annotations, and associated relevant metadata [7].

The audio data includes the mixture and, for a number of recordings, multi-track signals for all instrument sources except the tānpūrā. Since Carnatic music is primarily performed and enjoyed in a live performance setting, the audio recordings gathered for the Saraga dataset are all recorded in this context, and hence contain some leakage interference in the individual stem signals.

Alongside these audio recordings, Saraga provides automatically extracted annotations, such as the predominant pitch track of the vocalist's melody and rhythmic beats, and manual annotations, such as melodic patterns and musical sections. Finally, the dataset includes editorial metadata such as performer names, concert/composition titles, and musical tags such as melodic (rāga) and rhythmic (tāla) modes, which are crucial for this repertoire.

Whilst Saraga has proven to be a valuable resource for the analysis of IAM, there are nonetheless many challenges and important research questions for Carnatic music for which Saraga is insufficient. Some of these deficiencies – such as representativeness (e.g., instrument diversity, number of rāgas, demographics), completeness of annotations, and data access – have been pointed out in its open peer-review [8]. However, no new version of the dataset addressing said problems has been made available, and hence such deficiencies persist. Furthermore, Saraga contains automatically extracted features, which although may have been state-of-the-art at the time, could well be improved using more modern algorithms [9] and models [10].

In this work, we introduce *Saraga Audiovisual*, a new dataset built according to the principles of the original Saraga, that encompasses 42 new concerts totalling more than 60 hours of Carnatic music recordings. By including new artists, compositions, rāgas, and tālas, we improve the diversity and representativeness of the data. The new collection comprises multi-track audio, video recordings, and human pose estimation data, the latter two of which are entirely new modalities which are currently not considered in the first Saraga dataset. We hope that this multimodal data

will power further research of musically relevant problems in Carnatic music and encourage the development of underexplored research strands, particularly in music's visual and kinetic aspects [11–13]. We also improve documentation, access, and tools, considering the issues raised for Saraga [8], and provide a detailed description of the new dataset regarding musical metadata and coverage.

To showcase the value of the new dataset we present two benchmarking experiments to (a) demonstrate that the features extracted from the new audio-visual data are useful for the analysis of codependencies between performer body movement and vocalisations, an active research area in IAM analysis [14–21]; and (b) show that the novel multi-track audio is valuable for Music Source Separation (MSS) in Carnatic music, a low-level feature extraction task for which distributed pre-trained models in the literature do not generalize [22].

## 2. BACKGROUND AND RELATED WORK

Digital technology has brought new research methods to musicology [23, 24]. With digital archives and computer science techniques, researchers can study music corpora more systematically and quantitatively [25, 26]. Hence, creating appropriate datasets and research corpora for different music traditions is a fundamental concern in music information research (MIR) [27–32]. Computational musicological studies have used various data sources: scanned sheet music, symbolic scores, audio/video recordings, and motion capture data [11–13, 21, 33–36].

With few exceptions [36, 37], [1] almost all openly available datasets in the literature for Carnatic and Hindustani music are compiled from the IAM corpora in CompMusic [28], and more recently, from the Saraga dataset, for which multi-track audio recordings and manual and automatically extracted annotations are available [7].

Dataset distribution is a major concern in the music information research community [38], in which data plays a key role, especially given the advent of DL models. Saraga is currently accessed through Python notebooks, but the process is complex, not standardized, and hindered by bugs and dependency incompatibilities. Such a distribution method requires regular maintenance, which is expensive and time-consuming. A unified and functioning access point for the canonical version of the dataset, and a documented toolkit to browse through the recordings and annotations are not available.

One other important limitation of Saraga is that it contains only audio recordings. However, music is not only an auditory experience; multiple lines of research have demonstrated that the visual and kinetic aspects are all part of what music fundamentally is [39–41]. Thus, a comprehensive study of music performance requires auditory, visual and kinetic components [11].

In the case of Carnatic music, visual cues like hand/head gestures and performer gaze can provide the artists contextual information for an improved dynamic on stage, whilst also playing a more individualistic expressive role. This has been investigated in various IAM studies using motion capture data and pose estimation extracted from video [14–21]. For this reason, a dataset of Carnatic music should ideally include as much of this multimodality as possible, which we are enabling through the contribution of the video recordings in the proposed dataset.

[1] Using the IEMP North Indian Rāga: `https://osf.io/ks325/`, and Karnatak ālāpana multimodal dataset: `https://osf.io/6huvd/` respectively

## 3. DATASET DESCRIPTION

Saraga Audiovisual aims to address some of the aforementioned issues attributed to the first version of Saraga. In this section, we present the proposed improvements, which are mainly based on fixes, new recordings, and the novel visual modality. Although the new dataset falls entirely in the Carnatic repertoire, the proposed pipeline could be extended to Hindustani Music in the future.

### 3.1 New concerts

A total of 42 new concerts are released as part of Saraga Audiovisual, including multi-track audio and video for all concerts. The multi-track audio covers three main stems: vocals, violin, and mṛdaṅgam for all renditions, with the addition of ghaṭam and tānpūrā for 9 other concerts. The audios are all stereophonic and encoded at 44.1 kHz. Since the audio is recorded during live performances, the individual stems contain interference from the other sources.

These 42 concerts consist of a total of 235 individual performances of 223 unique compositions from 131 lead and accompanying artists. All performances include manually annotated section annotations. 10 distinct tālas and 113 distinct rāgas are represented, an increase of 55 on the existing Saraga dataset. Figure 1 shows the combined statistics for the case of the frequency of occurrence of the same rāga performances over Saraga and Saraga Audiovisual. Our aim is to increase the representation of existing rāgas whilst including unrepresented rāgas.

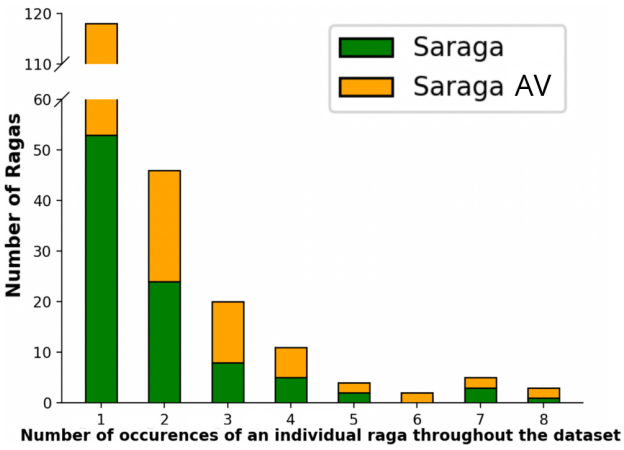


**Figure 1**. Number of occurrences of individual rāgas combining the two datasets. X-axis represents number of occurrences, whilst Y-axis indicates how many rāgas there are with $1, 2, ..., 8$ occurrences.

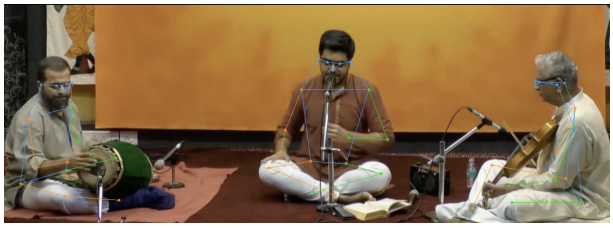

**Figure 2**. A video frame fragment from Saraga Audiovisual. The lead singer is VR Raghava Krishna, the violinist is VV Ravi and the mridangist is Guru Raghavendra.

| Content | Saraga | Saraga AV |
|---|---|---|
| Total number of recordings | 249 | 235 |
| Total number of artists | 64 | 131 |
| Number of compositions | 202 | 223 |
| Total number of rāgas | 110 | 235 |
| Unique rāgas in collection | 96 | 113 |
| Total number of tālas | 10 | 10 |
| Total duration of the dataset | 52.7 | 64.8 |

**Table 1**. Content comparison of the Carnatic subset of Saraga and the Saraga Audiovisual dataset (Saraga AV).

The most popular performance format today is vocal-led, either by a single or multiple vocalists. Despite the fair criticism of the shortage of instrumental recordings [8], we decide to consider only concerts led by a singer in Saraga Audiovisual. Moreover, the singing voice is extensively explored in the MIR literature, with numerous models designed to address various problems and research questions, offering opportunities for leveraging, training, and fine-tuning functional systems. We refine the statement around representativeness of Saraga to clarify that our dataset is intentionally vocal-centered.

### 3.2 Video recordings and human pose estimation

The videos corresponding to the concerts are rendered at 1080p and have a frame rate of 25 fps. They are recorded with a fixed wide-view position to frame all performing artists throughout the concert.

Figure 2 depicts the recording setting. The videos are recorded in a traditional concert set up with microphones occluding the view of the artists at most times. For example, if we observe a singer in this setting, they are in a seated position with the microphone directed towards their mouth. Consequently, the microphone head occludes the mouth, and the stand hampers the view of the singer's hands in several instances. In general, occlusions can hinder human pose estimation by a very large margin. After careful examination of several human pose estimation models, we choose MMPose [42], a DL model which performs extremely well, given the tricky setting. We extract human skeletons with 17 key points through its 2-D model. See Figure 3 for an example of the gesture estimation on a Saraga Audiovisual example video recording.

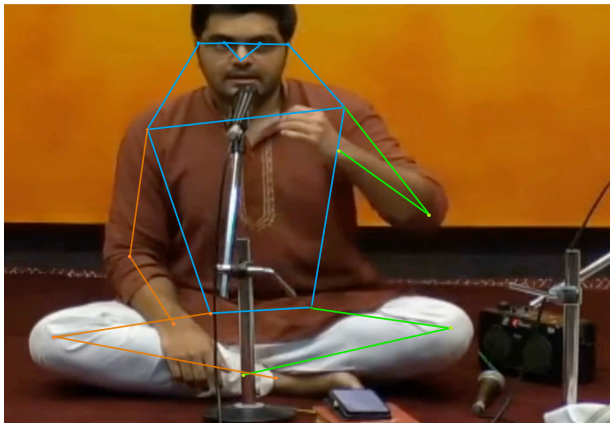

**Figure 3**. Gesture extraction with MMPose. The artist depicted in the figure is VR Raghava Krishna

### 3.3 Improving dataset access

Hassle-free and canonical access to the Saraga audio, annotations, and also metadata is an issue raised by the community [8]. We implement a `mirdata` [2] loader for Saraga Audiovisual to download, load, and browse through the canonical dataset and easily filter the data by musically important aspects such as rāga, tāla, artist, and tonic. These functions are also available through `compIAM` [3], where models and algorithms for the computational analysis of Carnatic music are also available.

### 3.4 Further improvements

Note that some of the features in Saraga are automatically extracted. Despite not being manually collected, these allow for faster, consistent, and reproducible research as we bypass the need to compute them multiple times. Since Saraga was first published, much research has been carried out by the MIR community, and new models to more reliably extract such features are available. Within the context of this work we compute the melody curves of the novel recordings using the Carnatic-optimized FTA-Net [10].

## 4. EXPERIMENTS

In this section we present two experiments using the audio and visual components of Saraga Audiovisual.

### 4.1 Multimodal study

There exists various studies that demonstrate the relationship between gesture and musical motifs in an IAM context [14–21]. Whilst most focus on the North Indian, Hindustani style, a recent study by Pearson et al. presents a quantitative attempt at characterising codependencies between the body movement and vocalisations of Carnatic performers using a combination of predominant pitch tracks extracted from audio, and motion tracking data captured using an inertial measurement system on the body during performance [37]. In an effort to demonstrate the value of

[2] https://github.com/mir-dataset-loaders/mirdata
[3] https://github.com/MTG/compIAM

| Performer | Rāga | Dur. |
|---|---|---|
| Ashwin Srikant | Siṁhēndramadhyamam | 09:17 |
| Raghava Krishna | Siṁhēndramadhyamam | 05:36 |
| Aditi Prahlad | Pūrvīkalyāṇi | 08:15 |
| Prithvi Harish | Pūrvīkalyāṇi | 08:01 |

**Table 2**. Saraga Audiovisual performances used for multimodal experiment in Section 4.1. Duration's (mm:ss) correspond to the rāga ālāpana section of the performance, the rest of the performance is not used for analysis.

Saraga Audiovisual in supporting such studies, we reproduce a part of Pearson et al's analysis here using data extracted from the proposed multimodal dataset. In the original study, performer gesture data is extracted using motion capture equipment, since here we rely on inferring this information from video, we make some changes to that part of the process, outlined in the following section. All other steps remain identical to the original study and we refer the reader to the paper for more details.

#### *4.1.1 Experimental setup*

We reproduce Pearson et al's Analysis 1: *Do sonic motif DTW distances covary with spatiotemporal patterns of gesture?* Our sonic data for such analysis is extracted from 4 performances (Table 2) in Saraga Audiovisual, from which we extract 4 time series corresponding to the rāga ālāpana section of the performance audio; (1) f0 – the predominant vocal melodic line, measured in cents above the performer tonic, extracted using a Carnatic-specific methodology [10], (2) $\Delta$f0 – the first derivative of f0, (3) loudness, $L$, computed as $L = 10 \cdot log_{10} \frac{S}{ref}$, where $S$ is the power spectrum of the raw audio signal and $ref$ is its maximum value, and (4) the spectral centroid of the raw audio signal. Our gestural data is extracted using MMPose and limited to the performer's left and right hands (see Section 3.1, and Figure 3), from which we compute the first and second derivatives to obtain two subsequent time series of velocity and acceleration, respectively, resulting in 6 gestural time series. The 6 time series are resampled so as to have identical sampling rates of 24 Hz, and all 10 time series are smoothed using a 2nd-order Savitzky-Golay filter with a window length of 125 ms.

In each of the 4 performances, we identify regions of repeated melodic motifs using a Carnatic-specific methodology [4]. For each motif, we isolate the corresponding segment in our 4 sonic and 6 gestural time series, and discard the gestural time series corresponding to the non-dominant hand. The dominant hand of the performer for each pattern is determined as that which has the highest kinetic energy, $K.E$, computed from the velocity tracks, $v$, where $K.E = \frac{mv^2}{2}$ and $m$ is the mass of the moving body part, assumed equal for both sides. We note that for over 98% of the identified motifs, the ratio in $K.E$ between the dominant and non-dominant hand is greater than 1.2, i.e. there is almost always a clear dominant hand. 70% of motifs are identified as left-handed and 30% as right-handed. The gesture space for each motif is transformed such that left and right-hand gestures occur in the same space by mirroring all right-handed gestures in the y-axis, and such that the gestural space origin corresponds to the centroid of the body of the performer. This centroid is determined for each motif as the centroid of the trapezoid corresponding to the performer's body, provided by MMPose and visible in Figure 3.2. The result is a selection of 269 non-overlapping motifs across the 4 performances, each represented by 4 sonic time series (f0, $\Delta$f0, loudness and spectral centroid) and 3 gestural time series (position, velocity and acceleration of the dominant hand).

For each pairwise combination of our 269 motifs, we compute the dynamic time warping (DTW) distance between each of their 6 sonic time series and 4 gestural time series, i.e. f0 compared to f0, hand position compared to hand position etc... Motifs are not compared to themselves and as such this constitutes 36,046 motif pairs. This analysis is concerned with whether there exists a relationship between the DTW distances of sonic and gestural features (sonic features: f0, $\Delta$f0, loudness and spectral centroid; gestural features: hand position, velocity and acceleration). For each combination of sonic to gestural features, we compute Spearman's rank correlation coefficient to quantify this relationship, both on a performer level and across all performers.

#### *4.1.2 Results*

The correlation analysis results are presented in Figure 4. Tests for which the p-value is greater than our significance level of 0.0001 are excluded and replaced with a grey square in the heatmap. It is not within the scope of this paper to discuss the results of this analysis in detail, nor do we consider the size of the data analysed sufficient to make any meaningful conclusions (0.5 hours of performance compared to 3.8 in the original study). We do, however, emphasize that even with this limited scope, we are able to identify significant relationships between sonic motif distances and spatiotemporal patterns of gesture using the Saraga Audiovisual dataset, corroborating the results of Pearson et al.'s study. As in that study, we show loudness as having the strongest correlation with gestural features across the performers and demonstrate how distinct the individual performer gesturing styles are, with great variation in the extent to which performers' gestures correlate with f0 and spectral centroid.

### 4.2 Fine-tuning MSS models with data with bleeding

The current state-of-the-art MSS models are based on DL architectures which are trained using multi-track recordings. Some models that are widely used, namely Spleeter [43] or Demucs [44], are trained with multi-track stems available through datasets like MUSDB18HQ [45] or MoisesDB [46], mainly including Western pop styles or related genres. Although many research works on the analysis of IAM use the available Spleeter model for source separation [13,47,48], these models do not generalize well for Carnatic music due to its varied instrumentation and

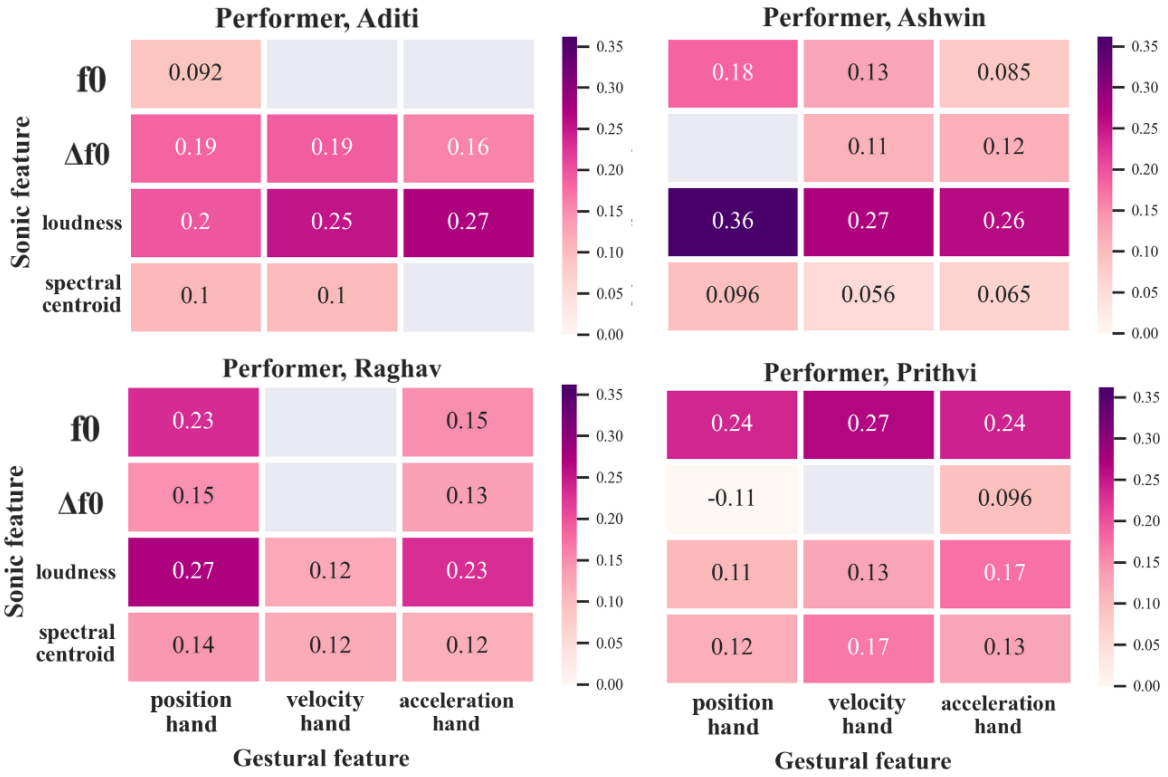


**Figure 4**. Spearman's rank correlation coefficient for all performers and on an individual level. Insignificant test results are represented by a grey square.

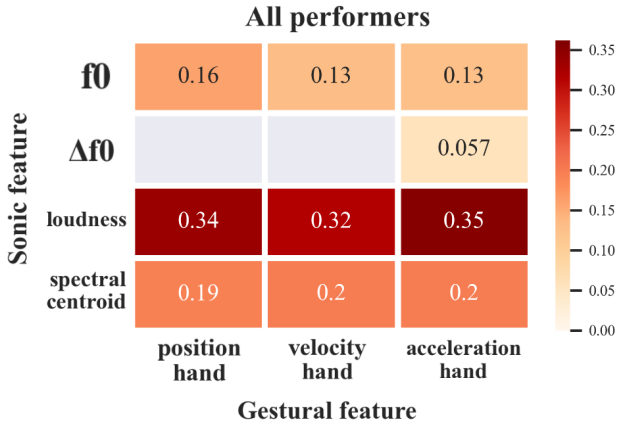


**Figure 5**. Spearman's rank correlation coefficient for all performers in the dataset. Insignificant test results are represented by a grey square.

idiosyncratic singing technique. In Carnatic music, the violin usually replicates or closely follows the melody of the singing voice. The tānpūrā provides an ambient canvas, and the mṛdaṅgam, a pitched percussion instrument, is strongly present. Existing MSS models are unfamiliar with such music styles and struggle to give a clean, separated singing voice stem.

There have been some efforts by the community to improve MSS for the use case of Carnatic music [22], since much research is done on top of vocals. Therefore the availability of isolated vocal recordings is highly valuable for the computational research of Carnatic music. Datasets like Saraga offer multi-track audio data, but given the fact that these recordings are from live concerts, there is leakage between sources that are part of the ensemble. The lack of clean multi-tracks for these music styles has been reported consistently, but Carnatic music is normally performed and recorded in a live setting. Therefore, recording musicians separately is not representative of how the tradition is generally performed. However, there has been some research on training MSS models with data with bleeding and some methods have been proposed to show the utility of data with bleeding [22].

Spleeter is a model based on a U-Net architecture that operates on the time-frequency domain. It is composed of a 6-layer encoder-decoder structure with skip-connections. Similar to most spectrogram-based separation models, Spleeter estimates $n$ separation masks that are multiplied by the input mixture spectrogram to separate the sources. The official implementation of Spleeter provides a framework to fine-tune the available pre-trained models in order to adapt the system to a specific domain [43].

In an ideal case, clean Carnatic multi-track stems would be essential to fine-tune Spleeter. However, we utilize the data with leakage that is part of the Saraga Audiovisual dataset in an attempt to set up a baseline for bespoke Carnatic vocal separation. We use the provided 2-stem Spleeter model, trained on a private dataset of 25k samples of 30s. We fine-tune Spleeter using Saraga and Saraga Audiovisual, aiming also to study the effect of the newly collected multi-track data. The models are fine-tuned for

600k steps with a constant learning rate of 1e-5. The fine-tuning process takes about a week in a TITAN XP GPU.

#### *4.2.1 Experimental setup*

Perceptual tests for MSS have gained interest in the research community, as objective metrics in [49] have been reported to not always correlate with the perceptual quality of MSS estimations [50]. Moreover, there is not a standardized and completely clean testing set for Carnatic separation. For that reason, we run a listening test with human subjects, including separations from recordings that we randomly collect from the Dunya dataset.

The listening test is based on the MUSHRA framework [51]. Subjects are asked to evaluate the vocal quality and the intrusiveness of other sources in separate stages. The scores are given on a scale from 1 to 5, with 5 being the maximum score. In each example, the subject is shown the original mixture as the reference stimuli, and the separations are shown unnamed and in a randomized order. The proposed subjective evaluation follows closely the ITU-T P.835 recommendation. We select and separate 6 Carnatic music concerts, ensuring diversity in audio quality and singer gender. Then, we randomly select a rendition from each concert, from which we collect a 30s chunk starting at a random point in time [22].

#### *4.2.2 Results*

We collect the results of the perceptual experiments and report the Mean Opinion Scores (MOS) per each model. We also report the Confidence Intervals (CIs) with $\alpha = 0.05$. A total of 20 subjects participated in the survey. Results are given in Table 3. While Spleeter samples are rated as having better vocal quality, Spleeter-FT-Sar improves over interference removal, while Spleeter-FT-SarAV is the most balanced solution among the three.

From the perceptual experiment, we conclude that Spleeter can better preserve the quality of the singing voice over the fine-tuned models. Using noisy data to fine-tune may be causing the model to lose some ability to properly discriminate the singing voice components. On the other hand, as the fine-tuned models improve on interference removal, we argue that it is possible for the pre-trained model to learn the instrumentation and vocal concepts of Carnatic music while preserving the knowledge to estimate separation masks for clean sources. In this particular experiment, Spleeter-FT-SarAV provides a balanced trade-off between artifacts and interferences. However, the overall performance is comparable to the other systems, suggesting that the multi-stem recordings have been obtained following the same peer-reviewed process in Saraga [7]. While establishing the baseline for Carnatic vocal separation, we also observe that leakage-aware systems such as [22] are still to be explored to take complete advantage of the multi-track data with leakage in both Saraga and Saraga Audiovisual, and outperform out of domain pre-trained models.

| | **Artifacts** | **Interferences** |
|---|---|---|
| **Spleeter** [43] | $3.89_{[3.75,4.04]}$ | $2.17_{[2.04,2.30]}$ |
| **Spleeter-FT-Sar** | $2.76_{[2.60,2.93]}$ | $3.80_{[3.68,3.93]}$ |
| **Spleeter-FT-SarAV** | $3.41_{[3.27,3.57]}$ | $2.88_{[2.74,3.02]}$ |

**Table 3**. MOS rating comparison between the default Spleeter [43] and a fine-tuned Spleeter using Saraga (FT-Sar) and Saraga Audiovisual (FT-SarAV). The higher, the better; 5 is the maximum rating. 95% CIs are also reported.

## 5. CONCLUSIONS

In this paper, we introduce Saraga Audiovisual, a multimodal dataset for the analysis of Indian Art Music, specifically of the Carnatic style. The dataset includes multi-track audio, fundamental frequency extractions from that audio, performance videos, and human pose estimation extracted from the video footage. The dataset also includes metadata like rāga, tāla, composition, and structural annotations like the ālāpana, kalpanā svara, niraval, and thani āvartana. The dataset is made available as a mirdata dataloader for easy and standardized access.

We perform two benchmarking experiments using the extracted audio and video features: (1) a multimodal analysis investigating the relationship between performer gesture and vocalisation, and (2) a fine-tuning experiment for Carnatic vocal source separation with audio data induced with leakage. Both experiments demonstrate the value of this data for music analysis in spite of imperfections in the automatically extracted feature data, such as audio leakage in the isolated instrument stems, or instability in the extracted pose estimations.

We expect Saraga Audiovisual to be a valuable resource for future work on tasks such as both vocal and instrument based leakage-aware source separation or predominant pitch extraction; and further multimodal studies of Carnatic music.

## 6. ETHICS STATEMENT

The Saraga Audiovisual dataset was recorded at the Arkay Convention Centre in Chennai. All of the artists appearing in the dataset gave informed consent for the dissemination of the data for research purposes through a consent form, and the Arkay Convention Centre was paid for their work in gathering the recordings.

With the release of this dataset, we wish to honour the cultural heritage of Carnatic music and safeguard its traditions. We recognize that to understand the distinctive intricacies of this culture and tradition, it is essential to go beyond computational methods alone. This understanding should not be oversimplified or broadly generalized.

## 7. ACKNOWLEDGEMENTS

This work was carried under the "IA y Música: Cátedra en Inteligencia Artificial y Música (TSI-100929-2023-1)",

funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA. Special thanks to Suhit Chiruthapudi for actively helping with the preparation of the dataset. We would like to acknowledge Arkay Convention Centre, Chennai and all the musicians included in the dataset. Special thanks to Aaditya Rangan Raghavan, Samiksha Sreekanthan, Adithya Srinivasan and R Sarang for their invaluable contributions to the dataset. We would also like to extend our thanks to Dr. Lara Pearson for her valuable insights on the paper. We would like to thank Marius Rodrigues for helping with the pose estimation of the videos and Serafin Schweinitz for his contribution to the dataset. Finally, we would like to acknowledge the 20 participants who agreed to undertake the perceptual test.

## 8. REFERENCES

[1] X. Serra, "A multicultural approach in music information research," in *Proc. of the 12th Int. Society for Music Information Retrieval Conf. (ISMIR),* Miami, USA, October 24-28 2011, pp. 151–156.

[2] E. B. Maria Panteli and S. Dixon, "A review of manual and computational approaches for the study of world music corpora," *Journal of New Music Research*, vol. 47, no. 2, pp. 176–189, 2018.

[3] 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 *Int. Symposium on Computer Music Multidisciplinary Research,* Tokyo, Japan, 2021, pp. 228–237.

[4] Nuttall, Thomas and Plaja-Roglans, Genís and Pearson, Lara and Serra, Xavier, "In search of sañcāras: tradition-informed repeated melodic pattern recognition in carnatic music," in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (ISMIR),* Bengaluru, India, 2022, pp. 337–344.

[5] T. Nuttall, X. Serra, and L. Pearson, "Svara-forms and coarticulation in carnatic music: an investigation using deep clustering," in *Proc. of the 11th International Conference on Digital Libraries for Musicology (DLFM),* Stellenbosch, South Africa, 2024, pp. 15–22.

[6] S. Paschalidou and I. Miliaresi, "Multimodal deep learning architecture for Hindustani raga classification," *Sensors and Transducers*, vol. 260, no. 2, pp. 77–86, 06 2023.

[7] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, "Saraga: Open datasets for research on indian art music," *Empirical Musicology Review*, vol. 16, no. 1, pp. 85–98, 2021.

[8] L. Pearson, "Cultural specificities in carnatic and hindustani music: Commentary on the saraga open dataset," *Empirical Musicology Review*, vol. 16, no. 1, pp. 166–171, 2021.

[9] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in essentia," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, 2020, pp. 266–270.

[10] G. Plaja-Roglans, T. Nuttall, L. Pearson, X. Serra, and M. Miron, "Repertoire-specific vocal pitch data generation for improved melodic analysis of carnatic music," *Transactions of the Int. Society for Music Information Retrieval*, vol. 6, no. 1, pp. 13–26, 2023.

[11] C. E. Cancino-Chacón and I. Pilkov, "The rach3 dataset: Towards data-driven analysis of piano performance rehearsal," in *Int. Conf. on Multimedia Modelling,* Amsterdam, The Netherlands, 2024, pp. 28–41.

[12] S. Nadkarni, S. Roychowdhury, P. Rao, and M. Clayton, "Exploring the correspondence of melodic contour with gesture in raga alap singing," in *Proc. of the 24th Conf. of the Int. Society for Music Information Retrieval (ISMIR),* Milano, Italy, 2023.

[13] M. Clayton, P. Rao, N. N. Shikarpur, S. Roychowdhury, and J. Li, "Raga classification from vocal performances using multimodal analysis." in *Proc. of the 23rd Int. Society for Music Information Retrieval (ISMIR),* Bengaluru, India, 2022, pp. 283–290.

[14] M. Rahaim, *Musicking bodies: Gesture and voice in Hindustani music*. Wesleyan University Press, 2013.

[15] M. Clayton, J. Li, A. Clarke, and M. Weinzierl, "Hindustani raga and singer classification using 2d and 3d pose estimation from video recordings," *Journal of New Music Research*, pp. 1–16, 2024.

[16] G. A. Fatone, M. Clayton, L. Leante, and M. Rahaim, "Imagery, melody and gesture in cross-cultural perspective," in *New perspectives on music and gesture*. Routledge, 2016, pp. 203–220.

[17] L. Leante, "The lotus and the king: imagery, gesture and meaning in a hindustani rāg," in *Ethnomusicology forum*, vol. 18, no. 2. Taylor & Francis, 2009, pp. 185–206.

[18] M. Charulatha, "Gesture in musical declamation: An intercultural approach," *Musicologist*, vol. 1, no. 1, pp. 6–31, 2017.

[19] P.-S. Paschalidou, "Effort in gestural interactions with imaginary objects in hindustani dhrupad vocal music," Ph.D. dissertation, Durham University, 2017.

[20] S. Paschalidou, T. Eerola, and M. Clayton, "Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical indian singing," in *Proc. of the 3rd Int. Symposium on Movement and Computing (SMC),* Thessaloniki, Greece, 2016, pp. 1–2.

[21] L. Pearson and W. Pouw, "Gesture–vocal coupling in karnatak music performance: A neuro–bodily distributed aesthetic entanglement," *Annals of the New York Academy of Sciences*, vol. 1515, no. 1, pp. 219–236, 2022.

[22] G. Plaja-Roglans, M. Miron, A. Shankar, and X. Serra, "Carnatic singing voice separation using cold diffusion on training data with bleeding," in *24th Int. Society for Music Information Retrieval Conf. (ISMIR),* Milano, Italy, 2023.

[23] E. Clarke and N. Cook, Eds., *Empirical musicology: Aims, methods, prospects*. Oxford University Press, 2004.

[24] T. Crawford and L. Gibson, Eds., *Modern methods for musicology: prospects, proposals, and realities*. Routledge, 2016.

[25] M. Müller, *Information retrieval for music and motion*. Springer Berlin, Heidelberg, 2007.

[26] D. Meredith, Ed., *Computational Music Analysis*. Springer Cham, 2016.

[27] R. Caro Repetto and X. Serra, "Creating a corpus of Jingju (beijing opera) music and possibilities for melodic analysis," in *Proc. of the 15th Conf. of the Int. Society for Music Information Retrieval (ISMIR),* Taipei, Taiwan, 2014.

[28] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Proc. of the Int. Computer Music Conf. (ICMC),* Athens, Greece, 2014.

[29] B. Uyar, H. S. Atli, S. Şentürk, B. Bozkurt, and X. Serra, "A corpus for computational research of Turkish Makam music," in *Proc. of the 1st Int. Workshop on Digital Libraries for Musicology (DLFM),* London, United Kingdom, 2014, pp. 1–7.

[30] M. Sordo, A. Chaachoo, and X. Serra, "Creating Corpora for Computational Research in Arab-Andalusian Music," in *Proc. of the 1st Int. Workshop on Digital Libraries for Musicology (DLFM),* London, United Kingdom, 2014, p. 1–3.

[31] R. C. Repetto, N. Pretto, A. Chaachoo, B. Bozkurt, and X. Serra, "An open corpus for the computational research of Arab-Andalusian music," in *Proc. of the 5th Int. Conf. on Digital Libraries for Musicology (DLFM),* Paris, France. New York, NY, USA: Association for Computing Machinery, 2018, p. 78–86. [Online]. Available: https://doi.org/10.1145/3273024.3273025

[32] M. Clayton, S. Tarsitani, R. Jankowsky, L. Jure, L. Leante, R. Polak, A. Poole, M. Rocamora, P. Alborno, A. Camurri, T. Eerola, N. Jacoby, and K. Jakubowski, "The interpersonal entrainment in music performance data collection," *Empirical Musicology Review*, vol. 16, no. 1, pp. 65–84, 2021.

[33] F. C. Moss, M. Neuwirth, D. Harasim, and M. Rohrmeier, "Statistical characteristics of tonal harmony: A corpus study of beethoven's string quartets," *PLoS One*, vol. 14, no. 6, p. e0217242, 2019.

[34] C. Weiß, M. Mauch, S. Dixon, and M. Müller, "Investigating style evolution of Western classical music: A computational approach," *Musicae Scientiae*, vol. 23, no. 4, pp. 486–507, 2019.

[35] H. Schreiber, C. Weiß, and M. Müller, "Local key estimation in classical music recordings: A cross-version study on schubert's winterreise," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, 2020, pp. 501–505.

[36] M. Clayton, K. Jakubowski, and T. Eerola, "Interpersonal entrainment in indian instrumental music performance: Synchronization and movement coordination relate to tempo, dynamics, metrical and cadential structure," *Musicae Scientiae*, vol. 23, no. 3, pp. 304–331, 2019.

[37] L. Pearson, T. Nuttall, and W. Pouw, "Landscapes of coarticulation: The co-structuring of gesture-vocal dynamics in karnatak music performance," 2024. [Online]. Available: osf.io/preprints/psyarxiv/npm96

[38] R. M. Bittner, M. Fuentes, D. Rubinstein, A. Jansson, K. Choi, and T. Kell, "mirdata: Software for reproducible usage of datasets," in *Int. Society for Music Information Retrieval Conf. (ISMIR),* Delft, The Netherlands, 2019.

[39] E. Clarke, "Meaning and the specification of motion in music," *Musicae Scientiae*, vol. 5, no. 2, pp. 213–234, 2001.

[40] R. I. Godøy, "Motor-mimetic music cognition," *Leonardo*, vol. 36, no. 4, pp. 317–319, 2003. [Online]. Available: http://www.jstor.org/stable/1577332

[41] Z. Eitan and R. Y. Granot, "How Music Moves: : Musical Parameters and Listeners Images of Motion," *Music Perception*, vol. 23, no. 3, pp. 221–248, 02 2006.

[42] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[43] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, pp. 1–4, 2020.

[44] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," 2019. [Online]. Available: http://arxiv.org/abs/1911.13254

[45] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," Dec. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[46] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR),* Milan, Italy, 2023.

[47] N. N. Shikarpur, A. Keskar, and P. Rao, "Computational analysis of melodic mode switching in raga performance." in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR),* Online, 2021, pp. 657–664.

[48] D. P. Shah, N. M. Jagtap, P. T. Talekar, and K. Gawande, "Raga recognition in Indian Classical Music using deep learning," in *Artificial Intelligence in Music, Sound, Art and Design: 10th Int. Conf. (EvoMUSART),* Sevilla, Spain, 2021, pp. 248–263.

[49] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," *Lecture Notes in Computer Science*, vol. 10891, pp. 293–305, 2018.

[50] E. Cano, D. Fitzgerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proc. of the 24th European Signal Processing Conf. (EUSIPCO),* Budapest, Hungary, 2016, pp. 1758–1762.

[51] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.