SANIDHA: A STUDIO QUALITY MULTI-MODAL DATASET FOR CARNATIC MUSIC

Venkatakrishnan Vaidyanathapuram Krishnan[1]   **Noel Alben**[1]   **Anish Nair**[1]
Nathaniel Condit-Schultz[1]

[1] School of Music, Georgia Institute of Technology, USA

`{vkrishnan65, noelalben3, anair323, natcs@gatech.edu}`

ABSTRACT

Music source separation demixes a piece of music into its individual sound sources (vocals, percussion, melodic instruments, etc.), a task with no simple mathematical solution. It requires deep learning methods involving training on large datasets of isolated music stems. The most commonly available datasets are made from commercial Western music, limiting the models' applications to non-Western genres like Carnatic music. Carnatic music is a live tradition, with the available multi-track recordings containing overlapping sounds and bleeds between the sources. This poses a challenge to commercially available source separation models like Spleeter and Hybrid Demucs. In this work, we introduce *Sanidha*, the first open-source novel dataset [1] for Carnatic music, offering studio-quality, multi-track recordings with minimal to no overlap or bleed. Along with the audio files, we provide high-definition videos of the artists' performances. Additionally, we fine-tuned Spleeter, one of the most commonly used source separation models, on our dataset and observed improved SDR performance compared to fine-tuning on a pre-existing Carnatic multi-track dataset. The outputs of the fine-tuned model with *Sanidha* are evaluated through a listening study.

1. INTRODUCTION

Carnatic music is a traditional "art music" genre from the Southern part of India. Carnatic Music is largely improvised, requiring all musicians to utilize a complex understanding of the melodic and rhythmic structures of the music to improvise coherently. Carnatic performances generally feature four to five musicians centered around a vocalist in the lead role. The core instruments are the violin, in both supportive and lead roles; the *mridangam*, a tonal two-sided drum that provides rhythmic support; and

the *ghatam*, a clay pot instrument that contributes rhythmic patterns to complement the *mridangam* in a higher frequency range. Carnatic Music performances are also accompanied by a *tanpura*, which constantly oscillates the *sa*, the tonic, and either the *pa*, the fifth or sometimes *ma*, the fourth. All the instruments are tuned to these frequencies, including the percussion instruments, which, too, have tonal qualities [1]. This leads to a significant overlap of frequency content, making Carnatic Music source separation almost impossible with simple dictionary learning methods [2].

Like most traditional music genres, Carnatic Music is performed live [1]. Thus, recordings of Carnatic Music lack multi-track isolation, as microphones inevitably capture signals from multiple instruments as well as the audience—these unwanted signals are known in music production as leakage or "bleed." This contrasts with Western pop music, where completely isolated multi-tracks are commonplace, and many source separation datasets are available [3–6]. The most extensive open-source Music Information Retrieval (MIR) dataset of Indian art music—the *Saraga* dataset [7]—exhibits significant leakage between different audio tracks: For example, the sound of the violin is audible in the vocal track - The bleeding of other sources into other microphones is significant [8–10].

1.1 Leakage Problem

Consider a signal $\mathbf{s}$, noise $\mathbf{n}$, and a mix $\mathbf{x}$, at 0 dB Signal-to-Noise Ratio (SNR): $\mathbf{x} = \mathbf{s} + \mathbf{n}$, where $\mathbf{x}, \mathbf{s}, \mathbf{n} \in \mathbb{R}^d$. Let $\mathbf{s_t}, \mathbf{n_t} \in \mathbb{R}^d$ such that they represent ground truth signal and noise with bleed. Assume no microphone sensor noise and no Room Impulse Response (RIR). Then

$$\mathbf{x} = \mathbf{s_t} + \mathbf{n_t} \tag{1}$$

$$\mathbf{s_t} = f(\mathbf{s}, \mathbf{n}) = \alpha\mathbf{s} + \beta\mathbf{n} \tag{2}$$

where $\alpha \in [0, 1]$ and $\beta \in [0, 1]$, using Eq. 1, it follows that

$$\mathbf{n_t} = g(\mathbf{s}, \mathbf{n}) = (1 - \alpha)\mathbf{s} + (1 - \beta)\mathbf{n} \tag{3}$$

Assume that functions f and g are linear time-invariant functions for all audios. However, the α and β values will vary for different signals in a general unclean dataset.

Let the source separation function trained with $(\mathbf{s_t}, \mathbf{n_t})$ as the ground truth be $\mathbf{F}$, such that

$$\mathbf{F}(\mathbf{x}) = (\hat{\mathbf{s}}, \hat{\mathbf{n}})$$
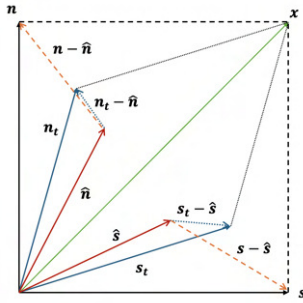

Figure 1. Problem of Poor Ground Truth

For simplicity, let us assume $\hat{\mathbf{s}}$, $\mathbf{s_t}$ and $\mathbf{n_t}$ lie in the same subspace as $\mathbf{s}$ and $\mathbf{n}$; $\mathbf{s}$ and $\mathbf{n}$ are orthogonal to each other i.e. $\mathbf{s}^T\mathbf{n} = 0$ as seen in Figure 1.

The most common metric used for evaluation and loss in the source separation community is the Signal-to-Distortion Ratio (SDR) and, more recently, the Scale-Invariant version of SDR called SI-SDR [11]. For simplicity, let's consider using SDR for evaluation since the idea can easily be extended to SI-SDR. SDR is defined by [12] for the `BSS_eval` toolbox (which is the same as classical SNR) as:

$$\text{SDR}_s = 10 \log_{10} \left(\frac{||\mathbf{s}||}{||\mathbf{s} - \hat{\mathbf{s}}||} \right)$$

Given that f and g functions vary for each audio, the SDR formula above is modified for data with bleed as:

$$\text{SDR}_{s,mod} = 10 \log_{10} \left(\frac{||\mathbf{s_t}||}{||\mathbf{s_t} - \hat{\mathbf{s}}||} \right)$$

These objective results from SDR, however good, will never truly represent what the original source must sound like. Training on data with a significant bleed will never push the predicted $\hat{\mathbf{s}}$ towards the actual source $\mathbf{s}$, since the loss function will be trained on the modified function dependent on sources with bleeding.

Furthermore, the result will be subpar after incorporating scale invariance [11]. If we calculate the norm of $\mathbf{s_t}$ and $\mathbf{n_t}$, using Eq. 2, 3, and the triangle inequality, we can prove:

$$||\mathbf{s_t}|| + ||\mathbf{n_t}|| \leq ||\mathbf{s}|| + ||\mathbf{n}|| \qquad (4)$$

This means that if we had to calculate the average SI-SDR of the signal and the noise with respect to the sources with bleed, the error would be significant. This error will be large when compared to calculating it with respect to "true" sources, which are inaccessible. It is also important to note that this was based on the assumption that all were in the same subspace, but that is never true in real scenarios, resulting in increased error.

Hence, the *Saraga* dataset cannot be used as accurate ground truth data for supervised source separation models for both training and especially evaluation, hindering the development of such models for Carnatic Music. As a workaround, some have attempted using source separation models like Spleeter [13], presumably trained on

a few or no Carnatic Music examples [9], directly on the vocal multi-track with bleeding for certain MIR tasks [8, 10]. However, attempts toward source separation for Carnatic using the currently available datasets have been made [2, 9, 14].

The stems obtained for Western Music datasets [3–6] are all from studio recordings, recorded separately and mixed, resulting in zero bleeds of other instruments in the multi-tracks. This allows for evaluation metrics such as SDR, Scale-Invariant SDR [11], Signal-to-Aritfacts Ratio, Signal-to-Interference Ratio (SIR), etc., to be used without problems. However, there is no such available dataset for Carnatic Music [2, 9], and since it is a live tradition, it is impossible to record the artists at separate times.

There have been a lot of datasets for Carnatic Music and Hindustani Music, which provide clean studio-quality data for individual instruments [15, 16]. However, there have been no completely isolated full live concert recordings of studio quality. To directly address this requirement, we present a new dataset of well-isolated multi-track recordings of Carnatic Music: *Sanidha*. The *Sanidha* dataset features audio and video recordings of Carnatic musicians playing together in real-time but in total isolation within a modern studio environment.

2. METHODOLOGY

Serra [17] proposed five essential considerations when creating new corpora: purpose, coverage, completeness, quality, and reusability. These considerations guided the creation of the *Saraga* dataset of Indian art music, [18], and we have worked to apply the same principles to the construction of *Sanidha*.

The isolated tracks for the commercial Western music source separation datasets are often created by the process of overdubbing in the studio. Carnatic Music must be improvised collectively in real-time, so parts cannot be "overdubbed" one at a time, thereby posing a significant challenge. Carnatic musicians listen closely to each others' playing and communicate extensively using visual cues. In particular, the vocalist often indicates the *taalam* (metric structure) with their hands. Visual cues are critical during fully improvisational sections like the *kalpana swaram* and *tani avartanam*. Consequently, the only way to record the music with audio isolation is for each musician to play in separate rooms while maintaining communication through audio and video.

2.1 Recording Sessions

We organized five Carnatic music concerts within our recording facility in March of 2024. Concert sessions lasted 2–3 hours, garnering an average of 1.6 hours of music per concert once silence between pieces was edited out. To perform these concerts, we recruited fifteen professional Carnatic musicians from Atlanta's thriving Carnatic music scene. All the artists voluntarily agreed to contribute to the dataset for research purposes, with no compensa-

tion[2].

Our musicians included three male vocalists, two female vocalists, four violinists, and six percussionists. Two out of five concerts featured a vocalist accompanied by the full set of core Carnatic instruments (violin, *mridangam*, and *ghatam*). The other three concerts proceeded without a *ghatam* player—which is not unusual for the style. Through the efforts of multiple talented musicians, we were able to capture gender diversity in the vocal timbre and a wide array of stylistic and improvisational approaches, which enhances the value of our data to the research community.

2.2 Recording Facility and Setup

The dataset was recorded in four rooms of the West Village Music Annex, in the Georgia Institute of Technology's campus in Atlanta, Georgia, USA. These rooms are multi-purpose spaces with large acoustic curtains, which enhanced our ability to control reverberation and maintain adequate isolation. The four isolated rooms have connection points wired to a single recording control room, including low-impedance, balanced analog audio, and digital video (SDI) connections. The control room uses a 32-channel digital mixing console to control audio routing and doubles as a multi-channel audio interface for digital audio recording into our Digital Audio Workstation (DAW). A *tanpura* drone, generated by a *shruti* box or a video from the internet, was also routed to each artist's headphones from the control room. We used the board's onboard reverb, compression, and equalization effects to create custom monitoring mixes sent to their headphones/inears, catering to individual artist needs and simulate the live traditional performing scenario of Carnatic Music.

Each artist's performance video was captured using a professional 4K video camcorder. The recorded video feed was then delivered through SDI cables from each room to the control room to generate a multi-source mixed feed, allowing us to transmit all four video feeds within a 2x2 grid (Figure 2). Musicians could see the 2x2 feed projected onto a screen in the performance room, allowing them to observe each other at all times.

Our musicians had little to no experience performing in a studio setting, isolated from each other, with headphones/in-ears on. Our efforts were focused on ensuring that the recording sessions were comfortable for the musicians and maintained the "natural" performance feeling as much as possible. Despite our best efforts, our musicians noted specific challenges performing within the constraints of the setup and sometimes felt that it slightly affected the quality of their performance.

Though our audio-monitoring setup achieved close to zero latency, we found that our video-monitoring setup lagged by about 50 ms, possibly due to the converters used to transmit the video feed to the projectors. This made it extremely difficult for the artists to coordinate with each other, since they could not follow the *taalam* or beat given by the vocalist. To overcome this problem, we used a *proxy-taalam* setup. One of our team members would sit in front of each artist (except the vocalist) and provide the visual *taalam* cue by focusing on just the audio feed from the vocalist. This setup was most helpful for our percussion artists; even the violinists appreciated it during the improvised *kalpana swaram* sections. The *proxy-taalam* setup allowed the musicians to play in time with each other, react to the cues from the vocalist similar to a live concert setup, and make the improvisational sections of Carnatic Music - *tani avartanam* and *kalpana swaram* sections possible.

We also identified a potential issue much later when we observed that some artists partially removed one side of their headphones in the middle of their performance. In some cases, artists required loud headphone output. This resulted in slight bleeding of the headphone output to the performer's microphone. To combat this, we shifted the monitoring system to in-ear monitors exclusively for all further concerts, which nullifies possible bleeding from headphones.

2.3 Audio Data

For each concert, we recorded six (excluding *ghatam*) or eight (full group) separate unprocessed audio tracks. Vocalists were recorded using a single microphone; the other instruments were recorded using two microphones each. We captured the violin and *mridangam* in a standard stereo (left-right) image. The *ghatam* recording setup used two microphones as well. A line-in track was used to record the *tanpura* drone.

In total, we have nearly eight hours of recorded music, across the five concert sessions. The recorded audio is in WAV format, with CD-standard sampling rate of 44.1 kHz and a bit depth of 16 bits. Table 1 displays all the individual concert durations.

2.3.1 Microphones

For each concert, we used different combinations of microphones, maximizing the sonic variety of the data. The choices of microphones were professional, studio-grade condenser microphones with cardioid polar pickup patterns, with each instrument requiring matched pairs of identical microphones. The use of high-fidelity condenser microphones contrasts with the dynamic microphones used commonly in traditional Carnatic music concerts. However, capturing the highest fidelity audio will produce the most broadly usable data. A series of non-linear operations can be performed at the post-processing stage to alter high-fidelity signals to sound more like dynamic microphones. The details of the microphones used for each instrument are stored in a JSON file located within respective concert folders.

For our first concert, the vocal microphone was placed close to the vocalist's mouth. We realized that this position obstructed the video of the performer's face. For all subsequent concerts, we corrected this by placing microphones

---

[2] The concerts were conducted with the approval of the Georgia Tech Institution Review Board (IRB) (ethics board), including two minors who were accompanied by their parents.

Concert	Instruments	Multi-tracks	Front-View Video	Side-View Video	Duration (hr)	Vocals Gender
1	Vocal	1	✓	-	1.08	Female
	Violin	2	✓	-		
	Mridangam	2	✓	-		
	Ghatam	2	✓	-		
2	Vocal	1	✓	✓	1.63	Male
	Violin	2	✓	-		
	Mridangam	2	✓	-		
3	Vocal	1	✓	✓	1.37	Male
	Violin	2	✓	-		
	Mridangam	2	✓	-		
	Ghatam	2	✓	-		
4	Vocal	1	✓	✓	1.97	Female
	Violin	2	✓	-		
	Mridangam	2	✓	-		
5	Vocal	1	✓	✓	1.92	Male
	Violin	2	✓	-		
	Mridangam	2	✓	-		

Table 1. Dataset Details

closer to chest level, pointing upwards towards the mouth, ensuring an obstruction-free video.

We placed microphones for the violin and *mridangam* on either side of the artist, at a distance of approximately 50 cm. This positioning ensured microphone stability, kept the video feed unobstructed, and highlighted each instrumentalist's gestures and hand movements. As the *ghatam* is a relatively quiet instrument, we placed the first microphone as close as possible to the playing surface. The second was pointed toward the opening of the *ghatam* at a distance of ≈ 30 cm. This can be seen in Figure 2.

2.4 Video Data

Performance video data for Carnatic is significantly limited compared to Hindustani music. The access to video data has given rise to a significant interest in the multi-modal analysis of Hindustani music among the MIR community [16, 19–21]. Our motivation to include video recordings with our dataset is to promote multi-modal research endeavors for Carnatic music.

All of our videos are recorded at 29.97 FPS in 1080p. The snapshot of the front view videos of each instrumentalist can be seen in Figure 2. The lighting for all the videos takes advantage of the many light sources available in the multi-purpose recording rooms.

For each concert, we successfully captured the front-view videos of every musician and included an additional side view of the vocalist. This combination is a first for a dataset of this kind.

The framing of the front-view videos is similar to the stills used in [20]. To ensure a solid background, we placed solid black sound panels behind the vocalists and solid yellow curtains behind the other artists, as seen in Figure 2.

Figure 2. Snapshot of the Front-view videos of Concert 3

2.5 Supplementary Information

2.5.1 Metadata

To fulfill Serra's completeness criteria, we collected annotations and metadata similar to *Saraga* [7]. This metadata is stored in separate JSON files for each song performed during the concerts. The metadata includes the composition name, original composer, and the performers' names and roles. We also include relevant music-theory information regarding the compositions, mentioning the rāgam, tālam, and song form.

2.5.2 Section Annotations

The song form is encoded as audio timestamps indicating the start and end of each major musical section for every song: the key sections are the *aalapana*, *pallavi*, *anupallavi*, *muktayi swaram*, *charanam*, *cittai swaram*, *kalpana swaram* and *neraval*. The performing musicians were consulted to review all of the metadata.

2.5.3 Pitch Annotations

Carnatic music contains two melody sources: the lead vocals and the violin, which complements the vocals. Since

we have clean vocals and violin data, the Melodia algorithm proposed by Salamon and Gòmez [22] was used to extract pitch (F0) contours for these two parts. The pitch tracks are stored in a two-column format, with the time stamps in the first column and the pitch values in the second.

2.5.4 Tonic Annotations

Obtaining the tonic frequency is relatively easy since we have a clean *tanpura* source within our multi-track data. We followed a similar approach used by Gulati et al. [23] and used Melodia [22] on the *tanpura* multi-track directly for the tonal feature extraction. The tonic does not change within a concert; hence, we included a single tonic file, which stores the tonic value in Hertz, inside each concert folder instead of having one for every song.

3. EXPERIMENTS

The experiments aim to cover the Coverage and Quality principles [17] introduced in Section 2 and demonstrate the value and usability of our new dataset with a simple source separation experiment.

It is important to note that our aim in this work is to demonstrate our data's potential through these preliminary experiments and not benchmark performance against the state-of-the-art results for source separation of Carnatic Music.

3.1 Experiment Setup

We ran a simple two-stem source separation fine-tuning experiment on *Sanidha* and *Saraga* datasets using the Spleeter model [13]. Two-stem Spleeter training requires the vocals, accompaniment, and mix audios. We fine-tuned the pre-trained model using three different approaches: (1) using the *Sanidha* dataset, (2) using the *Saraga* dataset, (3) using curriculum training [24, 25] by partly fine-tuning the model with *Saraga*, and then fine-tuning it further with the *Sanidha* dataset. The curriculum training strategy presents the data to the model in a meaningful order to learn better. Using these three models will help us evaluate the potential of our data and its performance when combined with other Carnatic Datasets, in this case, *Saraga*.

Since *Sanidha* has fewer concerts than *Saraga*, the major problem which could arise, is the possibility of overfitting. To potentially avoid this, the third model is fine-tuned on *Saraga* for 225K steps (90% of the total steps), while the rest 10% is finetuned on *Sanidha* for 25K steps.

3.2 Sanidha Data Preparation

Sanidha's audio data is of high quality as it was recorded in isolated spaces using condenser microphones with almost no bleed. Therefore, just linearly adding the signals to prepare mixes for training [26] will not be representative of the traditional Carnatic Concerts. To prevent this, we chose ten concerts from the *Saraga* dataset and used them as reference tracks to create two unique mixes for each of the five *Sanidha* concerts. Eight out of these ten *Saraga*

tracks are used as references for processing the training set and the remaining two are used for validation. The multiple mixes allow us to obtain twice the original amount of data. This can be considered as data augmentation since we have limited clean data. Our goal was to match the number of hours of training data used on the models individually trained on *Sanidha* and *Saraga* respectively, to make a fair comparison. The *Sanidha* training set makes up a total of 13.21 hours of audio data, and the validation set is 2.14 hours.

The critical mixing strategies for vocals and accompaniment include a combination of multiple non-linear and some time-varying operations - (1) Adding distortion, (2) Adding white noise, (3) Processing the stems through a digital amplifier plus cabinet models, (4) Heavy compression, (5) Adding reverb, (6) Attenuating the body of the instruments and vocals, and (7) High-cut filtering. Each of these operations is performed in varied amounts to match the sonic features of the reference tracks. The aim is to mix the tracks to emulate a real live concert while maintaining the isolated ground-truth audio.

The processed vocals ($\mathbf{v}$) and the processed accompaniment ($\mathbf{a}$) audios are linearly added at 0 dB SNR to create the mixture file ($\mathbf{m} = \mathbf{v} + \mathbf{a}$) for training. For the SNR computation, we consider the signal to be the vocals and the noise to be the accompaniment.

The third *Sanidha* concert was chosen for the validation set, as it has all of the typical instruments, including the *ghatam*. The rest of the concerts used in the training set maintain a good distribution of vocalist's gender and vocal timbre, as seen in Table 1. We made two unique mixes for each song in the validation set, which totaled to 2.14 hours of mixture audio data.

3.3 Saraga Data Preparation

Seven out of the eight references *Saraga* concerts described in Section 3.2 make up the training set for the *Saraga*-trained model. As *Saraga* consists of live multi-track recordings from Carnatic concerts, the accompaniment audio is created by linearly adding all the multi-track audios except the vocals. For the validation set of the *Saraga*-trained model, we selected the same reference concerts from *Saraga* that were used to create the mixes for the validation set in Section 3.2.

The ground truth multi-tracks used have an inherent bleed in them [8–10], as described in Section 1. The purpose of using a noisy validation set from *Saraga* is to evaluate the model purely trained on *Saraga*, assuming the *Sanidha* dataset never existed. However, the metrics obtained in Table 2 are on the validation set used for *Sanidha* training. The total training duration comes to 12.37 hours. The remaining unused concerts in *Saraga* are used for the perceptual tests.

4. EVALUATION

4.1 Objective evaluation

We compute the SDR, SIR, SAR, and also the SI-SDR of each of the models for the *Sanidha* validation set. Table 2

| Models	Hours	Source	*Sanidha* - Objective Evaluation				*Saraga* - Perceptual Evaluation	
			SDR	SIR	SAR	SI-SDR	Isolation	Audio Quality
Saraga	12.37	Vocals	7.66	17.05	8.02	6.65	0.596	0.627
		Accomp.	7.68	13.65	8.84	7.29	0.546	**0.532**
		Average	7.67	15.35	8.43	6.97	0.564	**0.580**
Sanidha	13.21	Vocals	**7.86**	**17.38**	**8.26**	**6.93**	0.598	**0.635**
		Accomp.	**7.87**	**13.96**	**8.99**	**7.52**	0.541	0.507
		Average	**7.87**	**15.67**	**8.63**	**7.22**	0.570	0.572
Mix	12.37 + 13.21	Vocals	7.63	16.88	8.00	6.62	**0.605**	0.621
		Accomp.	7.65	13.99	8.73	7.25	**0.561**	0.525
		Average	7.64	15.44	8.36	6.93	**0.583**	0.573

Table 2. Results
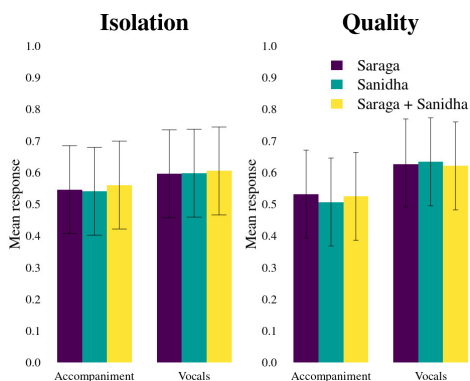
Figure 3. Mean participant responses across twelve conditions, with 95% confidence limits.

displays all the results.

We can see that the *Sanidha*-trained model has outperformed, however marginally, in all the objective metrics for vocals and the accompaniment separation. The improvement is only slight, perhaps because *Sanidha* was only trained on four concerts, while *Saraga* is trained on seven (which would mean seven unique vocalists as compared to four), even if the training hours are comparable. Also, the curriculum training technique performs almost similar to the *Saraga*-trained model.

4.2 Subjective evaluation

We conducted a listening study to evaluate the three source separation models and assess their perceptual effectiveness in isolating vocals and accompaniments in Carnatic Music.

The audio stimuli were selected after we randomly sampled four ten-second excerpts from four different *Saraga* recordings; If the randomly selected excerpt did not contain the three key instruments in Carnatic Music (vocals, violin, and the *mridangam*), we sampled again until an appropriate excerpt was identified. This iterative process ensured that our evaluation remains focused on relevant audio features while maintaining the unbiased nature of the sample selection.

In the listening study—approved by the Georgia Tech ethics board—fourteen participants listened to processed

versions of our selected excerpts. The survey was conducted in a manner similar to the MUSHRA framework [27]. All the participants responded to twelve questions for each excerpt, which focused on vocal isolation, vocal audio quality, accompaniment isolation, and accompaniment audio quality for the three models. This resulted in 48 questions per participant. These terms have been commonly used in subjective testing of source separation models [9, 28]. We used a slider-based metric for the evaluation, ranging from zero to one. Isolation and quality were explained with examples before the start of the survey and also presented as a reference for each question.

Average slider responses for the twelve conditions are shown in Figure 3 and in Table 2. We conducted a mixed-effects ANOVA on the data, with the participant and excerpt as random intercepts and the three variables (response type, target source, and model) as fixed effects. No effect was statistically significant, except for the target source (voice vs accompaniment), where participants tended to rate vocals higher in general ($\chi^2(8) = 45.97$, $p < .05$). This behavior is very similar to the objective results as well.

5. CONCLUSION AND FUTURE WORK

Although fine-tuning spleeter using *Sanidha* did not result in a significant source separation improvement, we cannot discount the importance of the availability of clean target sources for source separation. This is a clear distinction and advantage that our dataset collection methodology has over the existing *Saraga*. We can now use common metrics for source separation evaluation with a good degree of accuracy using our dataset, which was not possible with the existing *Saraga* dataset. Given the inherent challenges, our introduction of the *Sanidha* dataset marks a significant advancement in this domain. This novel dataset also presents an avenue for solving a multitude of other MIR and multimodal tasks in Carnatic Music.

We will soon expand our dataset and invite more musicians to conduct concerts using our methodology. With the resources at hand, we aim to promote computational analysis for Indian Art music and pave the path towards more accessible research resources within the community.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] A. Srinivasamurthy, S. Gulati, R. Caro Repetto, and X. Serra, "Getting started on computational musicology and music information research: an indian art music perspective," *Rao P, Murthy HA, Prasann SRM, editors. Indian art music: a computational perspective.[New Delhi]: Scheme for Promotion Academic and Research Collaboration, 2023. p. 3-38.*, 2023.

[2] J. Sebastian and H. A. Murthy, "Group delay based music source separation using deep recurrent neural networks," in *2016 International Conference on Signal Processing and Communications (SPCOM).* IEEE, 2016, pp. 1–5.

[3] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 2019.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[5] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *ISMIR*, vol. 14, 2014, pp. 155–160.

[6] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds. Cham: Springer International Publishing, 2017, pp. 323–332.

[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] 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 *International Symposium on Computer Music Multidisciplinary Research.* Springer, 2021, pp. 228–237.

[9] G. Plaja-Roglans, M. Miron, A. Shankar, and X. Serra, "Carnatic singing voice separation using cold diffusion on training data with bleeding," 2023.

[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 International Society for Music Information Retrieval*, Jun 2023.

[11] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[12] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[13] 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*, vol. 5, no. 50, p. 2154, 2020, deezer Research. [Online]. Available: https://doi.org/10.21105/joss.02154

[14] N. Dawalatabad, J. Sebastian, J. Kuriakose, C. C. Sekhar, S. Narayanan, and H. A. Murthy, "Front-end diarization for percussion separation in taniavartanam of carnatic music concerts," *arXiv preprint arXiv:2103.03215*, 2021.

[15] K. Subramani and P. Rao, "Carnatic violin dataset," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3940330

[16] M. Clayton, J. Li, A. R. Clarke, M. Weinzierl, L. Leante, and S. Tarsitani, "Hindustani raga and singer classification using pose estimation," 2021.

[17] X. Serra, "Creating research corpora for the computational study of music: the case of the compmusic project," in *AES 53rd International Conference: Semantic Audio; 2014 Jan 27-29; London, UK. New York: Audio Engineering Society; 2014. Article number 1-1 [9 p.].* Audio Engineering Society, 2014.

[18] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Georgaki A, Kouroupetroglou*

G, eds. Proceedings of the 2014 International Computer Music Conference, ICMC/SMC; 2014 Sept 14-20; Athens, Greece.[Michigan]: Michigan Publishing; 2014. Michigan Publishing, 2014.

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

[20] M. Clayton, P. Rao, N. N. Shikarpur, S. Roychowdhury, and J. Li, "Raga classification from vocal performances using multimodal analysis." in *ISMIR*, 2022, pp. 283–290.

[21] T. Kelkar, "Applications of gesture and spatial cognition in hindustani vocal music," Ph.D. dissertation, International Institute of Information Technology Hyderabad, India, 2015.

[22] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[23] S. Gulati, "A tonic identification approach for indian art music," *Unpublished master's thesis). Universitat Pompeu Fabra, Barcelona*, 2012.

[24] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[25] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[26] E. Manilow, P. Seetharman, and J. Salamon, *Open Source Tools & Data for Music Source Separation.* https://source-separation.github.io/tutorial, 2020. [Online]. Available: https://source-separation.github.io/tutorial

[27] 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, p. 8, 2018.

[28] A. Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.