



Repertoire-Specific Vocal Pitch Data Generation for Improved Melodic Analysis of Carnatic Music

DATASET

]u[ubiquity press

GENÍS PLAJA-ROGLANS **THOMAS NUTTALL** **LARA PEARSON** **XAVIER SERRA** **MARIUS MIRON**

*Author affiliations can be found in the back matter of this article

ABSTRACT

Deep Learning methods achieve state-of-the-art in many tasks, including vocal pitch extraction. However, these methods rely on the availability of pitch track annotations without errors, which are scarce and expensive to obtain for Carnatic Music. Here we identify the tradition-related challenges and propose tailored solutions to generate a novel, large, and open dataset, the Saraga-Carnatic-Melody-Synth (SCMS), comprising audio mixtures and time-aligned vocal pitch annotations. Through a cross-cultural evaluation leveraging this novel dataset, we show improvements in the performance of Deep Learning vocal pitch extraction methods on Indian Art Music recordings. Additional experiments show that the trained models outperform the currently used heuristic-based pitch extraction solutions for the computational melodic analysis of Carnatic Music and that this improvement leads to better results in the musicologically relevant task of repeated melodic pattern discovery when evaluated using expert annotations. The code and annotations are made available for reproducibility. The novel dataset and trained models are also integrated into the Python package `compIAM`¹ which allows them to be used out-of-the-box.

CORRESPONDING AUTHOR:

Genís Plaja-RoglansUniversitat Pompeu Fabra,
Barcelona, Spaingenis.plaja@upf.edu

KEYWORDS:

Carnatic Music; Data generation;
Vocal pitch extraction; Melodic
pattern discovery

TO CITE THIS ARTICLE:

Plaja-Roglans, G., Nuttall, T.,
Pearson, L., Serra, X., &
Miron, M. (2023). Repertoire-
Specific Vocal Pitch Data
Generation for Improved
Melodic Analysis of Carnatic
Music. *Transactions of the
International Society for
Music Information Retrieval*,
6(1), 13–26. DOI: [https://doi.
org/10.5334/tismir.137](https://doi.org/10.5334/tismir.137)

1 INTRODUCTION

Carnatic Music is a prominent art music tradition that originated in the royal courts and temples of South India and is still performed today in concert halls and at temple festivals. Alongside Hindustani Music (originating in North India), it constitutes one of the two main musical traditions of Indian Art Music (IAM). With its fan base encompassing millions of listeners and practitioners around the globe (Gulati et al., 2014), the interest in developing computational approaches for the analysis of Carnatic Music has grown in recent years (Tzanetakis, 2014). The most common contemporary instrumentation in Carnatic concerts involves a solo vocalist accompanied by violin and percussion, therefore, the vocal melody is highly musicologically significant in this style. In this work we contribute to analytical work on this musical repertoire by improving the automatic extraction of predominant vocal melody, or pitch, from mixed recordings using a Carnatic Music informed generation of ground-truth data for this task.

Because isolated vocal audio signals for Carnatic Music are scarce, the predominant melody is usually extracted from audio mixtures (Rao et al., 2014; Gulati et al., 2014, 2016; Ganguli et al., 2016; Nuttall et al., 2021). However, predominant pitch extraction is a difficult and not completely solved problem (Bittner et al., 2017). In particular, Carnatic Music constitutes a difficult case for vocal pitch extraction. Although performances place strong emphasis on a monophonic melodic line from the soloist singer, heterophonic melodic elements also occur, for example from the accompanying violinist who shadows the melody of the soloist often at a lag and with variation. Also, there is the tanpura (plucked lute that creates an oscillating drone) and pitched percussion instruments.

Many well-known predominant pitch extraction methods are heuristic-based (Rao and Rao, 2010; Durrieu et al., 2010; Salamon and Gomez, 2012). More recently, these have been outperformed by Deep Learning (DL) approaches (Kum et al., 2016; Kum and Nam, 2019; Yu et al., 2021). Data to train these models, however, is scarce and limited to Western music styles (LabROSA, 2005; Bittner et al., 2014). In this work, we hypothesize that these models generalize poorly to Carnatic Music given the melodic and instrumental uniqueness of this tradition, with domain transfer or re-training proving difficult due to the lack of Carnatic vocal pitch annotations (Benetos et al., 2018).

In order to understand why vocal pitch extraction is important for the computational analysis of Carnatic music we should first consider the melodic structure of the style. Rāgas are the primary melodic frameworks in Carnatic Music, best expressed through their characteristic sañcāras (melodic patterns, phrases or motifs) (Ishwar et al., 2013; Gulati et al., 2014). Although

rāgas are also conceptualized as having constituent pitch positions (svarasthānas) – which might be presented as something like a musical scale – in practice, the svaras (notes) are performed with gamakas (ornaments) that create melodic movement both on and between svaras (Krishna and Ishwar, 2012; Pearson, 2016). These can be experienced as small melodic atoms (Krishnaswamy, 2004; Morris, 2011). The gamakas that may be used on any given svāra are determined in part by the rāga, and therefore, also contribute to the identification of the rāga (Viswanathan, 1977). In fact, two rāgas may have the same constituent pitch positions, and in such cases it is the particular gamakas and sañcāras employed that disambiguate the two rāgas (Viswanathan, 1977; Kassebaum, 2000). Furthermore, the significance of sañcāras can be seen in the format known as rāga ālāpāna, in which the performer extemporises based on stock phrases: characteristic sañcāras that are typical of the rāga (Viswanathan, 1977; Pearson, 2021). As sañcāras and gamakas play an important role in rāga identity, ālāpāna performance and musical compositions, the automated discovery of such melodic patterns has formed an important strand in recent computational research on the style (Ishwar et al., 2013; Gulati et al., 2014; Nuttall et al., 2021).

Typically, the task of melodic pattern discovery in Carnatic Music has been based on time-series of predominant or vocal pitch extracted from audio recordings (Rao et al., 2014; Gulati et al., 2016; Nuttall et al., 2021). Approaches that depend on symbolic notation are another option, but manual transcription is extremely time consuming, and automated transcription is not a trivial problem in Carnatic music, since svaras are performed with gamakas that can radically shift the sound of the resulting unit from the theoretical pitch position referred to by the svāra name. Furthermore, the same svāra may be performed with different gamakas in different musical contexts, e.g. ascending or descending phrases (Ramanathan, 2004). Recently, the quantization of pitch tracks and identification of stable pitches and extremes (peaks and valleys) of the pitch contour has been used as an approach for creating a promising descriptive symbolic annotation/transcription of audio recordings (Ranjani et al., 2017, 2019). However, these rely on extracted pitch tracks and hence would benefit from any improvements in that process, such as those proposed here.

This work contributes to a more reliable and informative computational melodic analysis of Carnatic Music by improving the state-of-the-art for vocal pitch extraction in this tradition. The specific contributions are: (1) the Saraga-Carnatic-Melody-Synth dataset: a novel, large and open ground-truth vocal pitch dataset for Carnatic Music that is generated using a tradition-specific method inspired by Salamon et al. (2017), (2) we train a state-of-the-art data-driven vocal melody extraction

model (Yu et al., 2021) using the proposed dataset, and perform evaluation across multiple traditions, showing the positive impact of our tradition-specific approach for this task in a Carnatic context, and (3) we study the impact of the newly extracted pitch tracks on the musicologically relevant task of repeated melodic pattern discovery in Carnatic Music, evaluating the results using expert annotations.

2 LITERATURE REVIEW

2.1 DATASETS

The Dunya Carnatic and Hindustani corpora, built within the CompMusic project (Serra, 2014), provide many relevant datasets. Saraga (Srinivasamurthy et al., 2020) includes audio recordings of live performances containing, in many cases, close-microphone² recordings of violin, mridangam, ghatam and vocals. Although automatically-extracted pitch tracks are included in Saraga, these are not suitable for training data-driven models given the numerous errors that can result from automatic extraction using a heuristic algorithm (Salamon and Gomez, 2012), which may propagate to the trained models or prevent the training algorithm from reaching a decent minimum. To our best knowledge, no open ground-truth vocal melody dataset for Carnatic Music currently exists.

In general, pitch tracks are difficult and time consuming to annotate manually. Salamon et al. (2017) approached this issue by artificially creating these annotations in a reverse-engineering manner using an Analysis/Synthesis framework, first automatically extracting the pitch from melodic instruments, and then resynthesizing the audio signals over their corresponding extracted pitch tracks, forcing the harmonic structure of the regenerated signals to be built on top of frequency values at hand. Thus, the pitch tracks become ground-truth annotations for the resynthesized signals. However, this method is implemented for the MedleyDB dataset (Bittner et al., 2014), assuming isolated and studio-quality recordings of particular instrumentation and styles (mainly Pop and Rock), making it infeasible to be directly reproduced for the case of Carnatic music and its available datasets, which do not have the same characteristics.

2.2 METHODS FOR PITCH EXTRACTION

Melodia (Salamon and Gomez, 2012) is a heuristic-based algorithm for predominant melody extraction that has been broadly used for computational melodic analysis of Carnatic Music (Koduri et al., 2014; Gulati et al., 2014; Ganguli et al., 2016; Nuttall et al., 2021). Atli et al. (2014) adapt the heuristic rules in Melodia so that the longest detected pitch segments are prioritized, useful for musical contexts in which sustained melodic lines are recurrent. This approach is named *PredominantMelodyMakam*

(PMM). In addition to including 20 parameters that need to be manually set, Melodia is not optimized to discriminate melodic sources, which represents an important problem for Carnatic Music in which all instruments are pitched and the violin plays a prominent role.

Recent DL-based models for predominant pitch extraction can focus on a specific source (Kum et al., 2016; Kum and Nam, 2019; Yu et al., 2021). In this work, we refer to the Frequency-Time Attention Network (FTA-Net) (Yu et al., 2021), a deep neural network that achieves state-of-the-art performance for vocal pitch extraction. FTA-Net is formed by: (1) the *Frequency-Time Attention branch*, which consists of four layers, each formed by a Frequency-Temporal Attention (FTA) module connected to a Selective Fusion (SF) module. FTA modules mimic human hearing behaviour by drawing attention to the predominant melody source, while the SF dynamically selects and fuses the attention maps created by the FTA modules, (2) the *Melody Detection branch* is formed by four fully-stacked convolutional layers to perform iterative downsampling of the input data. In an ablation study, this branch is shown to improve the voicing detection.

FTA-Net is fed with the Combined Frequency and Periodicity (CFP) features (Su and Yang, 2015), which combine the power spectrum (frequency domain), the generalized cepstrum (time domain), and the generalized cepstrum of the spectrum (frequency domain), blending information of harmonics (frequency domain) and sub-harmonics (temporal domain) to emphasize the fundamental frequency and facilitate its detection. CFP are widely-used for predominant pitch extraction (Hsieh et al., 2019; Yu et al., 2021).

2.3 METHODS FOR MELODIC PATTERN DISCOVERY

A common pipeline for melodic pattern discovery in Carnatic music is to apply predominant pitch extraction to audio signals and compute pairwise distance measures between subsets of the data (Ishwar et al., 2013; Gulati et al., 2014; Rao et al., 2014; Nuttall et al., 2021). However, there is a lack of standardized baselines, evaluation datasets and metrics. Ishwar et al. (2013) and Gulati et al. (2014) involve experts after the results are collected to vote on their quality, whereas Rao et al. (2014) use expert annotations gathered beforehand. Owing to the expensive nature of creating expert annotations, they are limited in number. The Saraga dataset includes melodic pattern annotations, but for most recordings, only a few instances of a very limited number of patterns, normally from two to four, are available.

In this work we rely on Nuttall et al. (2021) for the melodic pattern finding approach. We refer the reader to the original paper and accompanying repository for a full explanation of the process, summarising here the particular elements relevant to the current paper. The

pipeline computes the matrix profile (Yeh et al., 2016), M , of the input pitch track, T , for a specified length, m . M returns, for each subsequence in T , the distance to its most similar subsequence in T . Similarity is computed using *non-z-normalised* Euclidean distance since the goal is to match subsequences identical in shape and y -location i.e. pitch. Pattern groups are identified by locating minima in the matrix profile to identify a *parent pattern* and querying the entirety of T with this parent. Groups include every retrieved subsequence with a distance to the parent below some threshold, ϕ , the only tunable parameter of the process.

In order to return full melodic motifs with plausible segmentation points, Nuttall et al. (2021) exclude subsequences that either contain long periods of silence (0 values in T) or stability (periods of constant pitch in T), since these are likely to lie at the borders of musically salient patterns (*sañcaras*) in Carnatic music.³ For that reason, here in this work we do not discard all stable pitches, as such points of stability are also salient in *rāga* performance (Viraraghavan et al., 2017). Instead, we have used long stability periods as cues for plausible segmentation, based on perceptual grouping tendencies in music perception (Deutsch, 1982; Deliege, 1987).

3 DATASET

3.1 DATA CREATION

Our approach to generate ground-truth vocal pitch annotations is inspired by Salamon et al. (2017). Our Carnatic-specific methodology intends to (1) characterize the main features of Carnatic melodies and instrumentation, and (2) adapt to the characteristics

of currently available Carnatic Music data. Related to the latter point, the input dataset for our process is the close-microphone Carnatic collection in Saraga (Srinivasamurthy et al., 2020). See Figure 1 for a complete diagram of the data generation process.

3.1.1 Leakage removal

The close-microphone audio in Saraga is recorded in live performances. The instruments are predominant in their corresponding tracks, but microphones also capture interference from the other instruments. We remove the accompaniment leaked in the singing voice using U-Net-based singing voice separation (Yu et al., 2021), which has recently shown promising results in source separation when the target source is predominant in the signal (Hennequin et al., 2020). We use the same approach to remove the voice from the recordings corresponding to the violin and mridangam. Finally, we compute the window-wise energy of the separated signals and remove leaked background noise with an energy below a predefined threshold. We do not aim at removing the mridangam interference in the violin signal and vice versa given that in the remixing step (Section 3.1.5) the accompaniment instruments are mixed together and all possible interferences between mridangam and violin are summed in the resulting accompaniment track.

3.1.2 Preliminary pitch extraction

We automatically extract the pitch curve from the cleaned vocal signal to be used as a reference for the singing voice resynthesis (Section 3.1.3). To account for possible unresolved interferences and source separation errors we decide against using a monophonic pitch tracker and use PMM (Atli et al., 2014) instead. In preliminary

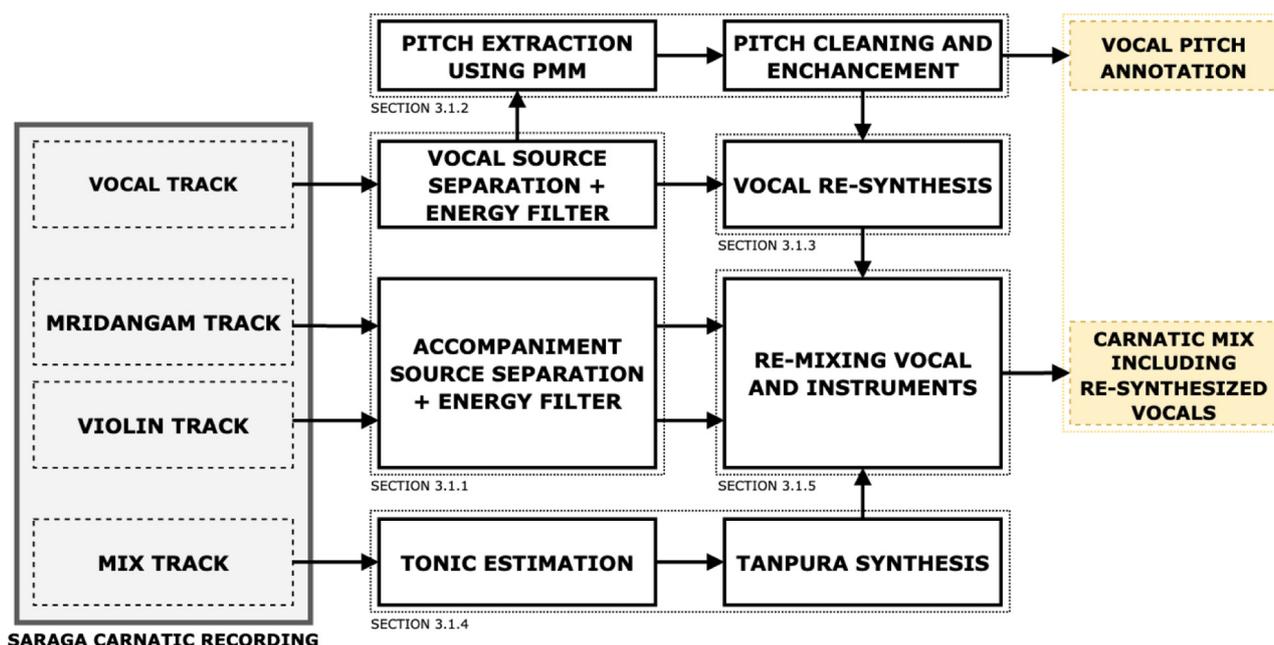


Figure 1 Block diagram of the data generation pipeline for a particular Saraga recording. We also indicate in which sections in this paper each building block is presented.

experiments, PMM showed better performance than the original Melodia algorithm in capturing gamakas, since the heuristics in PMM contribute to a better detection of sustained melodies. For the same reason, sporadic violin traces that may be present in the pitch are lessened at this point.

Next, the raw pitch track is enhanced through the following steps: (1) filling gaps shorter than 250 ms using 1D interpolation between boundaries (Ganguli et al., 2016; Nuttall et al., 2021), (2) smoothing the pitch curves using the Gaussian filter in Equation 1:

$$w(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} \quad (1)$$

where σ is the filter width and t the time index of the pitch track, and (3) restricting pitch track values to an interval of [80, 600] Hz (Venkataraman et al., 2020), annotating silent regions with 0 Hz.

Because the pitch extraction is run in an ideal scenario in which the singing voice is predominant and the leakage has been reduced, we can expect to obtain decent quality pitch tracks after post-processing. As a matter of fact, several computational studies of Carnatic Music build on top of pitch tracks obtained using a similar pipeline (excluding the leakage removal in Section 3.1.1) (Gulati et al. 2014, 2016; Ganguli et al., 2016), and the pitch tracks in Saraga are also computed as such (Srinivasamurthy et al., 2020). However, given the shortage of ground-truth vocal pitch annotations for Carnatic Music, we are currently unable to evaluate the quality of these objectively. Some available datasets have been validated by active listening (Eremenko et al., 2018), nonetheless, for the case of pitch track annotations, and especially in the melodically complex Carnatic context, such strategy may be time-consuming and a confident validation difficult to ensure. Another factor to consider is that this pipeline includes a combination of data-driven techniques and several heuristic steps, hence it is computationally expensive and its scalability may be compromised. Consequently, an improved, scalable, and straightforwardly runnable method would be beneficial.

3.1.3 Re-synthesis of the vocal audio signal

We denote the pitch track obtained with the pipeline in Sections 3.1.1 and 3.1.2 as *reference pitch*. Next, we reconstruct the spectral harmonic structure of the vocal signal on top of the reference pitch, so that it perfectly matches the new vocal audio signal and can be used for training and evaluating data-driven approaches. Thus, we consider the generated data as ground-truth. To perform the vocal resynthesis, we rely on the Harmonic plus Residual (HpR) model of Serra and Smith (1990).

We first compute the Short-Time Fourier Transform (STFT) of the original signal and identify the energy peaks in the magnitude spectrum and the corresponding values in the phase. The peaks are interpolated to infer

the frequency, magnitude and phase peak trajectories. More intuitively, for a vocal signal, these trajectories are the fundamental frequency plus the harmonics above, from which we also compute the magnitude and phase spectral values at each time-step. Next, we iterate through the trajectories while comparing the analyzed frequency peaks with perfectly formed harmonic series that are precomputed on top of the reference pitch. Given that both reference pitch and audio signal are sampled at the same rate, we can relate the time indexes in both. Let f_h be a certain precomputed harmonic over the reference pitch value, \hat{f}_h the closest peak to f_h in the trajectories obtained in the analysis, f_0 the reference pitch represented in Hz, and δ a harmonic deviation tolerance parameter. The lower δ is, the more restrictive we are. If the absolute difference between the analyzed frequency peak and the theoretical harmonic $|\hat{f}_h - f_h|$ is smaller than a pre-defined acceptance threshold β , we consider the analysis successful and accept the peak. As proposed in the Python implementation of Serra and Smith's HpR (Serra et al., 2015), a definition of β that is found to be effective is:

$$\beta = \frac{f_0}{3} + (\delta * \hat{f}_h) \quad (2)$$

We use β to filter out problematic regions that may lead to unnatural synthesis or artifacts. As per the successfully analyzed peaks, we substitute the analyzed frequency value in the trajectory by the perfect harmonic computed on top of the reference pitch track, to ensure the correspondence between the output regenerated signal and pitch annotation. We repeat the same approach for the entire number of harmonics in a window. When $|\hat{f}_h - f_h|$ is greater than β , we consider the synthesis not feasible and the peak is removed. If none or less than a predefined number of harmonics are accepted for a certain window, the window in the signal is completely silenced and the reference pitch value is set to 0. This prevention strategy averts unnatural synthesis and artifacts produced by the use of reference pitch values too distant from the actual note. Octave errors, apart from being reduced during pitch post-processing, are also addressed at this point, by setting to silence the affected regions. Reference pitch values with no corresponding predominant analyzed peaks are set to 0 Hz.

Next, we run additive synthesis using (1) the newly computed frequency peaks, and (2) the magnitude and phase trajectories originally analyzed from the input signal. The output of this operation is a resynthesized vocal signal whose harmonic structure has been generated on top of the reference pitch. We propose an adaptation of the HpR algorithm so that we are able to use the reference pitch at each window instead of automatically computing it using the Two-Way Mismatch algorithm (Maher and Beauchamp, 1994), as proposed by Serra et al. (2015). Moreover, the original

HpR performs the resynthesis over the frequency peaks that are estimated in the analysis of the input signal, while we mathematically compute the fundamental and harmonic frequency peaks over the reference pitch to ensure the annotation confidence.

Although we may miss the inharmonic content in the signal – mainly related to consonants and breathing – given that we resynthesize the sound using harmonic sinusoids, we focus on better characterizing ornamentations which are mainly present in sung vowels than consonants. Hence, we are interested in resynthesizing correctly the sung vowels while leaving the refinement of consonants for future work. As presented in Section 1, these ornamentations in Carnatic Music play a key role in the melodic framework and therefore are prioritized. By selecting a considerable number of harmonics we can output an intelligible vocal signal.

3.1.4 Tanpura synthesis

The tanpura instrument is not included in the close-microphone audio in Saraga but is an essential element in a Carnatic music arrangement, therefore we synthesize this instrument using a signal processing based model (Van Walstijn et al., 2016), which we adapt to allow any given tonic in Hertz as input. Since the tonic annotations in Saraga are automatically extracted and may be inconsistent for certain recordings, we estimate the tonic (Salamon et al., 2012) for an entire concert and use the most recurrent estimation, since a single tonic is expected to be maintained throughout the concert.

3.1.5 Remixing the stems

We generate an artificial mix by combining the synthesized vocal with the close-microphone audio tracks corresponding to the other instruments (tanpura, violin and mridangam). We use the remixing technique proposed by Bittner et al. (2014) for the MedleyDB dataset. The algorithm aims at estimating the mixing weights of the instrument audio signals using the original mix as reference by minimizing the following non-negative least squares objective constraining the mixing weights \mathbf{a} :

$$\|\mathbf{X}\mathbf{a}_i - \mathbf{Y}\|_2, \quad 5 > \mathbf{a}_i > 1 \quad (3)$$

Here, \mathbf{X} is a 3-rank tensor containing the STFT of the instruments and therefore shaped as (number of instruments, frequency bins, time steps), \mathbf{a} is the list of mixing weights, and \mathbf{Y} is the STFT of the original mix used as target, in this case shaped as (1, frequency bins, time steps). In contrast to Bittner et al. (2014), we include additional constraints to Equation 3 considering the characteristics of a Carnatic Music rendition. The vocal source is, without an exception, the predominant source in the performance and therefore, we restrict the vocal weights to be within [4, 5]. The tanpura is never predominant but a background generator very rich in

timbre, therefore we restrict its weights to [1, 3]. The final restriction is to avoid the violin to be louder than the vocal source, to ensure the predominance of such from the lead melodic sources. These rules contribute to the generation of a more natural Carnatic recording following the traditional mixing configuration for each instrument in the arrangement. We relate the remixed stems including the resynthesized vocal signal, with the corresponding pitch annotation we used as reference for the resynthesis.

3.2 DATA GENERATION SETUP

The presented pipeline is applied to the recordings in Saraga that have available close-microphone audio, 168 out of the total 249. There is no convention in the literature about the length in seconds of individual samples. Given that our dataset is intended to serve as training data (rather than to be listened to) and that the duration of Carnatic Music performances often last over an hour, we split the recordings into chunks of 30 s. The pipeline is run at a sampling rate of 44.1 kHz.

(Section 3.1.1) The source separation is performed using Spleeter (Hennequin et al., 2020). The threshold in the energy filtering for leakage removal is 1.25% of the peak energy in the signal.

(Section 3.1.2) PPM is applied using a window of 2048 samples and hop size of 128. Before running the algorithm, we apply an equal loudness filter to the signal (Salamon and Gomez, 2012). The pitch curve is Gaussian smoothed with a sigma of 1.

(Section 3.1.3) The window and hop sizes are set to 2048 and 128 respectively corresponding to the pitch extraction parameters. We use 30 harmonics, $\delta = 0.001$, and set to 5 harmonics the minimum to perform resynthesis. The remaining parameters can be found in the implementation referenced in Section 7.

(Section 3.1.4) The tanpura synthesis model is used out-of-the-box (including our modification presented in Section 3.1.4). Since the model is designed to synthesize a single pass throughout the instrument strings, the generated tone has a duration of ≈ 3 s. We generate multiple instances, concatenating them using a triangular window to obtain a 30 s signal.

(Section 3.1.5) We apply a single scalar mixing weight per instrument for the entire 30 s chunk. Given the extensive duration of Carnatic renditions and the sections in a performance, and also the roughly stable mix in live concerts, we perceptually observe no notable mixing imbalance in the audio chunks.

3.3 DATASET SPECIFICATIONS

The dataset resulting from the presented process is named Saraga-Carnatic-Melody-Synth (SCMS), and is made publicly available for research purposes (see Section 7). The dataset occupies 25GB of space. We split the dataset into train (11 artists, 1683 files, ≈ 845 minutes) and test

DATASET	CONSIDERED GENRES	LENGTH	% VOCAL	NO. SAMPLES	SAMPLE LENGTH	AVAILABLE AUDIO?
MedleyDB V1 [^]	Rock, pop, jazz, rap	≈447 min	57%	108	~20–600 sec	Upon request
MedleyDB V2 [^]	Rock, pop, jazz, rap	≈750 min	57%	196	~20–600 sec	Upon request
MDB-mel-synth [×]	Rock, pop, jazz, rap	≈190 min	64%	65	~20–600 sec	Yes
MIR1K [†]	Chinese pop	≈113 min	100%	1000	~4–13 sec	Yes
RWC [*]	Japanese & US pop	≈407 min	100%	100	~240 sec	No
ADC2004 [®]	Rock, pop, opera	≈10 min	60%	20	~30 sec	Yes
MIREX05 [®]	Rock, pop	≈6 min	80%	12	~30 sec	Yes
MIREX09	Chinese pop	≈167 min	100%	374	~20–40 sec	No
INDIAN08	Hindustani Music	≈8 min	100%	8	~60 sec	No
SCMS	Carnatic Music	≈1235 min	100%	2460	~30 sec	Yes

Table 1 Specification comparison between our SCMS and different state-of-the-art-melody datasets (Goto et al., 2002)^{*}, (LabROSA, 2005)[®], (Hsu and Jang, 2010)[†], (Bittner et al., 2014)[^], (Salamon et al., 2017)[×]. Table inspired by a similar comparison by Bittner et al. (2014).

(5 artists, 777 files, ≈390 minutes) sets, ensuring that the singer gender and tonic range are equally-distributed between both sets, and the recordings of a single artist are never in both train and test sets simultaneously. Audio chunks containing solely unvoiced melody segments are disregarded.

The audio files are in WAVE format, 16 bit depth, and sampled at 44.1 kHz. The melodic annotations are in CSV format with two columns: timestamp in seconds (hop size of 29 ms) and fundamental frequency values, represented in Hertz (Hz). Unvoiced timestamps are annotated with 0 Hz. The dataset includes a folder for the audio recordings, and a folder for the annotations. Note that audio and annotations for a particular chunk, artist, or even tonic, can be retrieved using the included JSON dataset metadata. As such, one can easily parse our proposed and curated splits to reproduce our results. See Table 1 for a detailed comparison of the specifications in terms of style and size between SCMS and available pitch extraction datasets.

4 EXPERIMENTS

4.1 VOCAL PITCH EXTRACTION

In order to evaluate the impact of the SCMS on the computational melodic analysis of Carnatic Music, we perform two vocal pitch extraction experiments using the generated data.

Experiment A – Cross-cultural evaluation of FTA-Net: Empirical comparison between FTA-Net trained and evaluated on data collections from both IAM and Western music domains. We aim at studying the extent of the domain-drift (Quiñero-Candela et al., 2009) between different musical repertoires in a DL-based vocal pitch extraction context.

Experiment B – Comparison between FTA-Net and Melodia: Empirical comparison between FTA-Net trained with the SCMS and baseline Melodia.

4.1.1 Experimental setup

The CFP features to feed FTA-Net are computed at a sampling rate of 8 kHz to reduce the computational expense. We use a window size of 2048 samples (256 ms at 8 kHz) and hop of 80 (10 ms at 8 kHz). We do not use the complete SCMS in our experimentation to prevent the difference in training data size to be a determining factor when comparing the models. Correspondingly, for each artist in the training set of the SCMS, we randomly select 85 samples, in total, from all renditions performed by said artist. Therefore, we consider $85 \times 11 = 935$ samples, from a total of 95 renditions performed by the 11 artists in training set. For training we use the ADAM optimizer with a learning rate of 0.001, batch-size of 16, and binary cross-entropy loss.

We denote the FTA-Net trained on the vocal subset of MedleyDB as FTA-W (W stands for Western Music). We resynthesize the singing voice in MedleyDB using our method in Section 3.1.3 and remix it back with the rest of instruments to bypass biases that may arise from the synthesis algorithm. We denote by FTA-C (C stands for Carnatic Music) the FTA-Net trained on the SCMS. We evaluate FTA-W and FTA-C on (1) the test set of SCMS, (2) a collection of Hindustani data of the same size, referred to as SHMS, and generated using our synthesis method applied only on selected mixture recordings where the singing voice is clearly dominant, since no close-microphone data is available in this case, (3) Western datasets from the literature – ADC2004, and MIREX05. We use the melody metrics of Bittner and Bosch (2019) including Voicing Recall (VR), Voicing False Alarm (VFA), Raw Pitch Accuracy (RPA), Raw Chroma Accuracy (RCA) and Overall Accuracy (OA).

4.2 MELODIC PATTERN DISCOVERY

Ultimately, vocal pitch tracks are used as features for further computational musicology research. To assess the impact of our extracted pitch tracks, we use them

for the musicologically relevant task of melodic pattern discovery. Nuttall et al. (2021) present a computational approach that utilizes the main melodic vocal line, extracted using Melodia, to identify and group melodic patterns in the sung melody from a performance in Saraga Carnatic. With the subsequent availability of expert annotations of melodic patterns in this performance (Section 4.2.2) we are able to apply this pattern finding approach to our newly extracted pitch tracks and the current state-of-the-art, Melodia, evaluating the results empirically.

Experiment C – Melodic pattern discovery comparison: We assess the suitability of eight pitch track candidates for melodic pattern discovery. To this end, a match between a detected pattern and our expert-annotated sañcāra and phrases is evaluated in terms of precision, recall, and F_1 -score.

4.2.1 Experimental setup

We extract eight pitch tracks from the audio of a performance from an artist not present in the training data of the SCMS: the Akkarai Sisters performance of a composition titled Koti Janmani,⁴ by the composer Oottukkadu Venkata Kavi, set in the Carnatic *rāga Ritigauḷa*, used also by Nuttall et al. (2021). Each of the eight pitch tracks correspond to one of the following four methods applied to either the audio of the mixed recording or the close-microphone vocal stem in Saraga:

1. **Melodia** – Current baseline method.
2. **Melodia-S** – Melodia applied to vocal source separated audio (Hennequin et al., 2020).
3. **FTA-W** – FTA-Net trained on MedleyDB, vocal stem passed through our resynthesis algorithm.
4. **FTA-C** – FTA-Net trained on the SCMS

For each of the eight pitch tracks we apply the melodic pattern discovery process of Nuttall et al. (2021). To account for the variable length of patterns, the process is run for each integer length in the range [2, 7]s. Pattern groups are restricted to a maximum of 20 per group, with a maximum of 20 groups for each pattern length. In practice, no pattern groups reach these limits. ϕ is determined individually for each pitch track and chosen as the threshold which delivers the maximum F_1 score when evaluated on our expert annotations (Table 4). We optimize for F_1 so as to control the quantity and relevance of the retrieved patterns.

It is important to consider that (1) there is some variation possible in where an annotator might choose to segment a longer pattern, and (2) the process only accepts fixed values of pattern length m rather than a range; consequently, since the annotations themselves are of variable length, the process will almost always return a pattern whose length does not correspond exactly to the annotation. Therefore, we consider a

returned pattern, R , to be a match with an annotation, A , if the intersection between them is more than two-thirds the length of A and more than two-thirds the length of R .

4.2.2 Melodic pattern annotations

As mentioned in Section 2, no open and complete annotations to evaluate the task of melodic pattern discovery are available. For the evaluation of the melodic pattern finding experiment we use expert annotations created, as part of this work, by a professional Carnatic vocalist. These contain all sañcāras in the audio recording, annotated in the software ELAN (Sloetjes and Wittenburg, 2008), using sargam syllables (the notation traditionally used by Carnatic musicians to refer to the underlying svaras). Sañcāras (musically meaningful phrases and motifs) are defined according to the annotator’s lived experience as a professional performer of 21 years standing. It should be noted that there exists no definitive list of such sañcāras, and although there will be a good degree of agreement between professional musicians, there will also be subtle points of difference in defining the borders of a sañcāra/phrase (the segmentation), as is also the case in the annotation of other musical styles (Bruderer, 2008; Nieto, 2015). Although these annotations are subjective to some degree, they have the virtue of being informed by expert knowledge of the style rather than based on an externally imposed metric, and thus should be highly relevant to Carnatic performance practice.

As sañcāras in Carnatic compositions are often varied on their subsequent repetitions, but are still recognizably connected musically, we capture these connections by grouping such variations together with an ‘underlying sañcāra’ annotation, where the underlying annotation is the first occurrence of the sañcāra group (typically the simplest version). Furthermore, as Carnatic music can be segmented hierarchically, wherein sometimes two or more shorter segments will lie within one longer segment, we also create longer phrase-level annotations which consist of either one long or several short sañcāras. This is done to capture the multi-level nature of musical structure (McFee et al., 2017; Popescu et al., 2021). Finally, the evaluation is made on the two levels – ‘underlying sañcāras’ and ‘underlying full phrases’.

5 RESULTS AND DISCUSSION

Experiment A. Cross-cultural evaluation of FTA-Net.

In Table 2 a cross-cultural evaluation of FTA-Net is displayed. We observe a notable performance difference due to the change of repertoire in the testing data. On the test set of SCMS, FTA-W performs worse than FTA-C,

≈20% less in terms of overall accuracy, showing that training the FTA-Net with data that does not include the idiosyncratic melodic features of Carnatic Music produces subpar performance on the Carnatic repertoire. We observe a similar trend when evaluating on the Western music datasets. We study a selection of Carnatic vocal pitch tracks extracted using FTA-W to identify specific problems in comparison to Western recordings. The most common observed problem is the effect of the violin, which typically overlaps with the lead singer, playing a similar melody line, causing the false alarm detection to rise (see VFA column in Table 2). Another observed problem is that FTA-W fails to capture the *gamakas* in detail. Such strong and fast vocal ornaments are not commonly seen in Pop/Rock.

Note the reduced number of false alarm detections of the FTA-C on the SCMS testing set, in addition to the improvement on pitch and chroma accuracies. The high performance that FTA-C achieves in the Carnatic repertoire is probably due to the fact that the SCMS is less diverse in terms of instrument arrangement and vocal style, representing as such the current performance practice in Carnatic music, where concerts with vocalists as soloists greatly outnumber those where other instruments act as the soloist. In Table 2 we also observe that the FTA-C generalizes better to Hindustani than to Western music, yet achieving better performance than the model trained on modern Western music, FTA-W, from which we can infer that the model is not overfitting to the SCMS but learning the idiosyncratic features of the music repertoire. Despite being two different music traditions, Carnatic and Hindustani include many common concepts and features.

Experiment B. Comparison between FTA-C and Melodia. In Table 3 we observe the comparison between FTA-C and Melodia. From this table we can conclude that (1) FTA-C is able to outperform Melodia, a broadly used pitch extraction algorithm for the computational analysis of Carnatic Music and (2) Melodia scores a vocal pitch extraction accuracy comparable to the performance reported in the original paper for diverse test datasets (Salamon and Gomez, 2012), which may suggest that the SCMS dataset includes standard quality audio with high-accuracy annotations.

Experiment C. Quantitative evaluation of melodic pattern discovery. Table 4 presents the recall, precision and F_1 score of the melodic pattern discovery algorithm applied to the eight pitch tracks. The results extracted from the FTA-C pitch tracks outperform the others in almost all metrics except for precision on the vocal stem, in which Melodia-S, which is the second best performing pitch track, is leading. However, FTA-C pitch track achieves +28.4% recall and +13.8% F_1 over Melodia-S for the mixed stem, and an improvement of +9.5% recall and +1.5% F_1 for the vocal stem.

We observe that Melodia and Melodia-S struggle to achieve anywhere near competitive results when applied to the mixed recording, but do provide comparable results for the vocal stems, however data of this kind is rarely available in practice. The results also suggest that using Spleeter to clean the accompaniment is not sufficient to solve the issue, showing the relevance of FTA-C. The “coverage” in Table 4, refers to the proportion of the returned pitch track which corresponds to non-zero values; note that this does not reflect the proportion of the pitch track

MELODY EXTRACTION METRICS

↓ TEST SET/MODEL →	VR		VFA		RPA		RCA		OA	
	FTA-C	FTA-W								
SCMS (test)	96.35	83.26	8.38	31.43	90.17	69.30	90.46	70.62	90.99	67.72
SHMS	91.25	80.18	17.04	17.53	78.96	68.76	81.78	70.20	81.39	73.84
MIREX05	86.74	89.21	21.40	19.23	68.11	73.94	69.68	74.18	72.44	76.66
ADC2004	77.25	87.79	29.17	27.94	64.01	77.98	66.62	79.98	64.46	77.32

Table 2 Performance comparison between FTA-Net trained using the SCMS (FTA-C) and MDB-synth (FTA-W). Results presented as percentages (%).

MELODY EXTRACTION METRICS

MODEL → ↓ TEST SET	VR		VFA		RPA		RCA		OA	
	FTA-C	MELODIA	FTA-C	MELODIA	FTA-C	MELODIA	FTA-C	MELODIA	FTA-C	MELODIA
SCMS (test)	96.35	85.75	8.38	17.17	90.17	77.51	90.46	79.81	90.99	77.07

Table 3 Performance comparison between FTA-Net trained using the SCMS (FTA-C) and Melodia (Salamon and Gomez, 2012). Results presented as percentages (%).

PITCH TRACK	STEM	COVERAGE (%)	PRECISION	RECALL	F1	NO. PATTERNS	NO. GROUPS	ϕ
Melodia	Mix	69.0	0.323	0.297	0.310	164	21	2.7
Melodia-S	Mix	71.4	0.341	0.371	0.356	170	20	2.8
FTA-W	Mix	74.8	0.250	0.007	0.113	4	2	2.9
FTA-C	Mix	80.3	0.396	0.655	0.494	283	66	2.2
Melodia	Vocal	76.0	0.514	0.574	0.542	181	48	1.0
Melodia-S	Vocal	75.3	0.523	0.574	0.547	197	50	1.0
FTA-W	Vocal	75.3	0.395	0.155	0.223	43	20	2.9
FTA-C	Vocal	78.0	0.485	0.669	0.562	227	49	2.4

Table 4 Comparison of different pitch extraction methods for melodic pattern discovery.

which is annotated *correctly*. The pitch tracks returned from FTA-C have slightly more coverage with a +8.9% and +3.3% improvement on Melodia-S for the mix and vocal stem respectively. Note also that the pitch tracks extracted using FTA-W (trained on Western music) perform considerably worse than all others in metrics, coverage and number of groups returned.

There are 148 annotated patterns in total; 41 are identified by both FTA-C and Melodia-S. FTA-C identifies a further 56 that Melodia-S is unable to and Melodia-S identifies a further 14 that FTA-C is unable to. In total there are 37 patterns that neither manage to identify. Figure 2 illustrates four examples of annotations that only FTA-C is able to identify. It is obvious from exploring these comparative plots that Melodia struggles to correctly annotate regions of quite intense oscillations from the mixed recording, as seen in the third plot in Figure 2 between ~153.0 s and ~153.6 s, again between ~154.0 s and ~154.3 s, and further throughout the annotation plots that can be found in the accompanying GitHub repository.

Figure 3 shows four instances of motif group 39 starting at different parts of the performance. The pitch track in the Figure has been extracted by FTA-C on the mixed recording. We can see two forms of the same phrase – the first, appearing at 3:01 and 3:19 is the simplest version (covering the svara annotation “nnsndmgnmns”), while a melodic variation on this phrase can be seen at 6:57 and 7:08. In fact, an underlying sañcāra included in this phrase occurs eight times in this recording, seven of which were found and correctly placed into one group of related sañcāras, notwithstanding the fact they show considerable variation. Such returned groups of motifs could be of interest to musicologists who wish to examine variations in performance of sañcāras and related phrases across audio recordings, for example, for the purposes of comparative and/or historical performance analysis, as well as the analysis of musical compositions. Any such study would ultimately depend on the quality and accuracy of their extracted vocal pitch annotations.

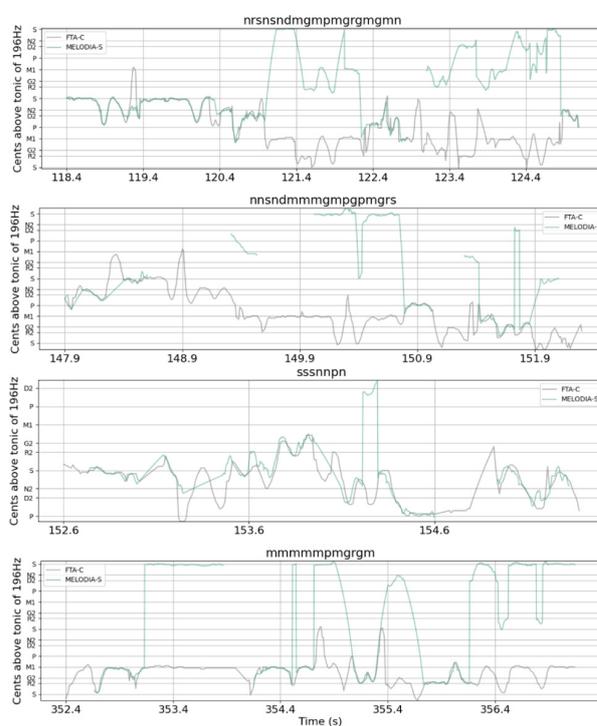


Figure 2 Four different example patterns identified by FTA-C but disregarded by Melodia-S (pitch extraction run on the mixture audio).

The reader is referred to the accompanying repository where **Experiment C** is run for two additional recordings including different artists and rāgas.

6 CONCLUSIONS

In this work we present a methodology to automatically generate a novel, large, and open collection of vocal pitch annotations for Carnatic Music: the SCMS. We use the Saraga dataset which serves as input for a bespoke Analysis/Synthesis method that accounts for the features and challenges of Carnatic Music, as well as data availability for this style. To study the impact of the SCMS, we then train a state-of-the-art vocal pitch extraction model, aiming to equal the performance that

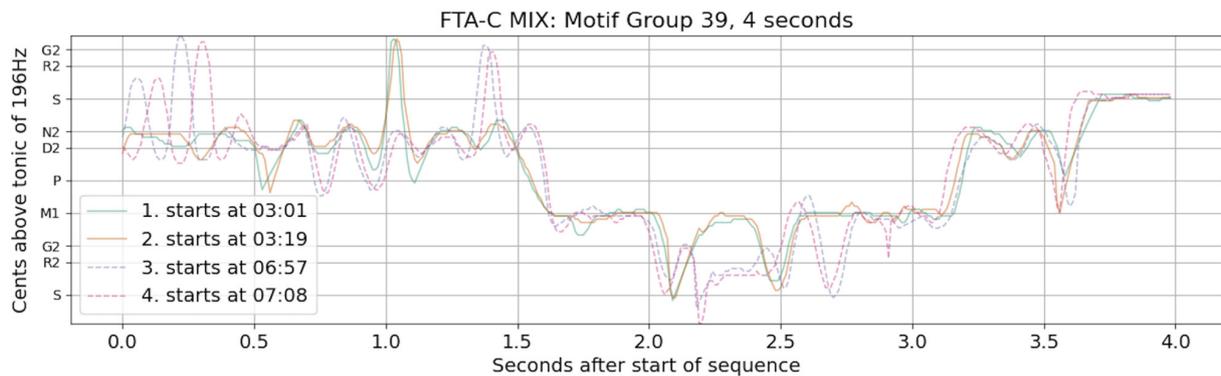


Figure 3 4 occurrences of motif 39 retrieved using FTA-C on the mixed recording. The dashed and solid lines refer to two distinct variations of the same underlying melodic pattern.

this model obtains for currently available datasets, which mainly include Pop, Rock, and related music repertoires. We run a comprehensive cross-cultural evaluation, comparing two models trained on Carnatic and Western music respectively, across Carnatic, Hindustani, and Western test sets. We show that the model trained on the SCMS is able to obtain state-of-the-art vocal pitch extraction results for Carnatic Music. We provide proof and discussion of the importance of the SCMS dataset by running melodic pattern discovery experiments using the improved vocal pitch tracks. The results show that our newly extracted pitch tracks boost the performance on discovering expert annotated melodic patterns, while Melodia is a viable approach only if separated vocal recordings are available. We also note that FTA-Net trained on the available datasets for pitch extraction prior to this work performs very badly and reduces the discovery of Carnatic melodic patterns. From this, we observe there is a need for repertoire-specific data to break the glass-ceiling for the task of vocal melody extraction and subsequent musicologically-relevant computational analysis across a wider range of musical styles from different cultural contexts. In future work we aim at updating the data generation method with latest repertoire-specific technologies, targeting a cleaner version of the SCMS. We are also interested in properly extending the proposed methodology to the Hindustani tradition and if possible, to other repertoires.

7 REPRODUCIBILITY

We publish the SCMS in Zenodo (DOI: 10.5281/zenodo.5553925). The implementations for (1) data generation, (2) training FTA-Net with the SCMS, in addition to the pre-trained models for Carnatic Music, are available here: <https://github.com/MTG/carnatic-pitch-patterns>. Also, we include code to run the experiments and visualizations in the paper. The expert pattern annotations are made available in the repository. Both SCMS and FTA-Carnatic have been integrated into `compIAM` to be used out-of-the-box.

NOTES

- 1 <https://github.com/MTG/compIAM>.
- 2 In this work we use *multi-track* to refer to completely separated instrument recordings, and *close-microphone* for a multi-track scenario with leakage (typically recorded live or in the same room).
- 3 We studied the sañcara annotations created for this work and found that long, stable regions tend to lie either outside or at the borders of annotations.
- 4 <https://musicbrainz.org/recording/5fa0bcfd-c71e-4d6f-940e-0cef6fbc2a32>.

ACKNOWLEDGEMENTS

We gratefully acknowledge the significant contribution to this paper made by the professional Carnatic vocalist and musicologist Brindha Manickavasakan who created the sañcara annotations used in one of the evaluation procedures.

FUNDING INFORMATION

This work is funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI) within the Musical AI Project–PID2019-111403GB-I00/AEI/10.13039/501100011033.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Genís Plaja-Roglans  orcid.org/0000-0003-3450-3194
Universitat Pompeu Fabra, Barcelona, Spain

Thomas Nuttall  orcid.org/0000-0001-6316-1424
Universitat Pompeu Fabra, Barcelona, Spain

Lara Pearson  orcid.org/0000-0002-5073-8738
Max Planck Institute for Empirical Aesthetics,
Frankfurt, Germany

Xavier Serra  orcid.org/0000-0003-1395-2345

Universitat Pompeu Fabra, Barcelona, Spain

Marius Miron  orcid.org/0000-0002-2563-075X

Universitat Pompeu Fabra, Barcelona, Spain

REFERENCES

- Atli, H., Uyar, B., Şentürk, S., Bozkurt, B., and Serra, X.** (2014). Audio feature extraction for exploring Turkish Makam music. In *Proceedings of the International Conference on Audio Technologies for Music and Media (ATMM)*, Ankara, Turkey.
- Benetos, E., Dixon, S., Duan, Z., and Ewert, S.** (2018). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30. DOI: <https://doi.org/10.1109/MSP.2018.2869928>
- Bittner, R., and Bosch, J.** (2019). Generalized metrics for single-F0 estimation evaluation. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 738–745.
- Bittner, R., McFee, B., Salamon, J., Li, P., and Bello, J.** (2017). Deep salience representations for F0 estimation in polyphonic music. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 63–70.
- Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J.** (2014). MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 155–160.
- Bruderer, M.** (2008). *Perception and Modeling of Segment Boundaries in Popular Music*. PhD thesis, J.F. Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven, The Netherlands.
- Deliege, I.** (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's grouping preference rules. *Music Perception*, 4(4):325–359. DOI: <https://doi.org/10.2307/40285378>
- Deutsch, D.** (1982). Grouping mechanisms in music. In *The Psychology of Music*, pages 99–134. Academic Press. DOI: <https://doi.org/10.1016/B978-0-12-213562-0.50008-5>
- Durrieu, J., Richard, G., David, B., and Fevotte, C.** (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, pages 564–575. DOI: <https://doi.org/10.1109/TASL.2010.2041114>
- Eremenko, V., Demirel, E., Bozkurt, B., and Serra, X.** (2018). Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 483–490.
- Ganguli, K. K., Gulati, S., Serra, X., and Rao, P.** (2016). Data-driven exploration of melodic structures in Hindustani music. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 605–611.
- Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R.** (2002). RWC Music Database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pages 13–17.
- Gulati, S., Serrà, J., Ishwar, V., and Serra, X.** (2014). Mining melodic patterns in large audio collections of Indian Art Music. In *Proceedings of the International Conference on Signal Image Technology and Internet Based Systems*, pages 264–271. DOI: <https://doi.org/10.1109/SITIS.2014.73>
- Gulati, S., Serrà, J., Ishwar, V., and Serra, X.** (2016). Discovering raga motifs by characterizing communities in networks of melodic patterns. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. DOI: <https://doi.org/10.1109/ICASSP.2016.7471682>
- Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M.** (2020). Spleeter: A fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, pages 1–4. DOI: <https://doi.org/10.21105/joss.02154>
- Hsieh, T., Su, L., and Yang, Y.** (2019). A streamlined Encoder/Decoder architecture for melody extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160. DOI: <https://doi.org/10.1109/ICASSP.2019.8682389>
- Hsu, C., and Jang, J.** (2010). On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, pages 310–319. DOI: <https://doi.org/10.1109/TASL.2009.2026503>
- Ishwar, V., Dutta, S., Bellur, A., and Murthy, H. A.** (2013). Motif spotting in an alapana in Carnatic music. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 499–504.
- Kassebaum, G. R.** (2000). Karnatak raga. In Arnold, A., editor, *The Garland Encyclopaedia of World Music*, pages 89–109. Garland, New York. DOI: <https://doi.org/10.1080/09298215.2013.866145>
- Koduri, G. K., Ishwar, V., Serrà, J., and Serra, X.** (2014). Intonation analysis of ragas in Carnatic music. *Journal of New Music Research*, pages 73–94.
- Krishna, T. M., and Ishwar, V.** (2012). Carnatic music: Svara, gamaka, motif and raga identity. In Serra, X., Rao, P., Murthy, H., and Bozkurt, B., editors, *Proceedings of the 2nd CompMusic Workshop*, pages 12–18.
- Krishnaswamy, A.** (2004). Melodic atoms for transcribing Carnatic music. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*.
- Kum, S., and Nam, J.** (2019). Joint detection and classification of singing voice melody using convolutional recurrent neural networks. *Applied Sciences*, 9(7). DOI: <https://doi.org/10.3390/app9071324>

- Kum, S., Oh, C., and Nam, J.** (2016). Melody extraction on vocal segments using multi-column deep neural networks. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 819–825.
- LabROSA.** (2005). MIREX05 and ADC2004. Retrieved September 1, 2021 from <https://labrosa.ee.columbia.edu/projects/melody/>.
- Maher, R., and Beauchamp, J.** (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of The Acoustical Society of America*, 95. DOI: <https://doi.org/10.1121/1.408685>
- McFee, B., Nieto, O., Farbood, M. M., and Bello, J. P.** (2017). Evaluating hierarchical structure in music annotations. *Frontiers in Psychology*, 8:1337. DOI: <https://doi.org/10.3389/fpsyg.2017.01337>
- Morris, R.** (2011). Tana varnams: An entry into raga delineation in Carnatic music. *Analytical Approaches to World Music*, 1(1):1–27.
- Nieto, O.** (2015). *Discovering Structure in Music: Automatic Approaches and Perceptual Evaluations*. PhD thesis, New York University.
- Nuttall, T., Plaja-Roglans, G., Pearson, L., and Serra, X.** (2021). The matrix profile for motif discovery in audio—an example application in Carnatic music. In *Proceedings of the 15th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 109–118.
- Pearson, L.** (2016). Coarticulation and gesture: An analysis of melodic movement in South Indian raga performance. *Music Analysis*, 35(3):280–313. DOI: <https://doi.org/10.1111/musa.12071>
- Pearson, L.** (2021). “Improvisation” in play: A view through South Indian music practices. In *The Routledge Handbook of Philosophy and Improvisation in the Arts*, pages 446–461. Routledge, Abingdon. DOI: <https://doi.org/10.4324/9781003179443-35>
- Popescu, T., Widdess, R., and Rohrmeier, M.** (2021). Western listeners detect boundary hierarchy in Indian music: A segmentation study. *Scientific Reports*, 11(1):1–14. DOI: <https://doi.org/10.1038/s41598-021-82629-y>
- Quinoñero-Candela, J., Sugiyama, M., Lawrence, N., and Schwaighofer, A.** (2009). *Dataset Shift in Machine Learning*. MIT Press. DOI: <https://doi.org/10.7551/mitpress/9780262170055.001.0001>
- Ramanathan, N.** (2004). Sargam and musical conception in Karnataka system. In *Sargam as a Musical Material*. Dr. Prabha Atre Foundation, Pu. La. Deshpande Maharashtra Kala Academy, Mumbai.
- Ranjani, H. G., Paramashivan, D., and Sreenivas, T. V.** (2017). Quantized melodic contours in Indian Art Music perception: Application to transcription. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, (ISMIR)*, pages 174–180.
- Ranjani, H. G., Srinivasamurthy, A., Paramashivan, D., and Sreenivas, T. V.** (2019). A compact pitch and time representation for melodic contours in Indian Art Music. *The Journal of the Acoustical Society of America*, 145(1):597–603. DOI: <https://doi.org/10.1121/1.5087277>
- Rao, P., Ross, J. C., Ganguli, K. K., Pandit, V., Ishwar, V., Bellur, A., and Murthy, H. A.** (2014). Classification of melodic motifs in raga music with timeseries matching. *Journal of New Music Research*, 43(1):115–131.
- Rao, V., and Rao, P.** (2010). Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 18(8):2145–2154. DOI: <https://doi.org/10.1080/09298215.2013.873470>
- Salamon, J., Bittner, R., Bonada, J., Bosch, J., Gómez, E., and Bello, J.** (2017). An analysis/synthesis framework for automatic F0 annotation of multitrack datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 71–78. DOI: <https://doi.org/10.1109/TASL.2010.2042124>
- Salamon, J., and Gomez, E.** (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, pages 1759–1770. DOI: <https://doi.org/10.1109/TASL.2012.2188515>
- Salamon, J., Gulati, S., and Serra, X.** (2012). A multipitch approach to tonic identification in Indian Classical music. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 499–504.
- Serra, X.** (2014). Creating research corpora for the computational study of music: The case of the CompMusic project. In *Audio Engineering Society International Conference*, pages 1–9.
- Serra, X., Serra, F., and Gulati, S.** (2015). sms-tools. GitHub (<https://github.com/MTG/sms-tools>).
- Serra, X., and Smith, J.** (1990). Spectral modeling synthesis: A sound analysis/synthesis based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14:12–24. DOI: <https://doi.org/10.2307/3680788>
- Sloetjes, H., and Wittenburg, P.** (2008). Annotation by category—ELAN and ISO DCR. In *Proceedings of the 6th international Conference on Language Resources and Evaluation (LREC)*.
- Srinivasamurthy, A., Gulati, S., Repetto, R., and Serra, X.** (2020). Saraga: Open datasets for research on Indian Art Music. *Empirical Musicology Review*. DOI: <https://doi.org/10.18061/emr.v16i1.7641>
- Su, L., and Yang, Y.** (2015). Combining spectral and temporal representations for multipitch estimation of polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, pages 1600–1612. DOI: <https://doi.org/10.1109/TASLP.2015.2442411>
- Tzanetakis, G.** (2014). Computational ethnomusicology: A music information retrieval perspective. In *Proceedings of the 40th International Computer Music Conference*, pages 112–117.
- Van Walstijn, M., Bridges, J., and Mehes, S.** (2016). A real-time synthesis oriented tanpura model. In *Proceedings of*

the 19th International Conference on Digital Audio Effects (DAFx), pages 175–182.

Venkataraman, M., Boominathan, P., and Nallamuthu,

A. (2020). Frequency range measures in Carnatic singers. *Journal of Voice*. DOI: <https://doi.org/10.1016/j.jvoice.2020.08.022>

Viraraghavan, V. S., Aravind, R., and Murthy, H. A. (2017).

A statistical analysis of gamakas in Carnatic music. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, (ISMIR)*, pages 243–249.

Viswanathan, T. (1977). The analysis of raga ālapana in South Indian music. *Asian Music*, 9(1):13–71. DOI: <https://doi.org/10.2307/833817>

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.

A., Silva, D. F., Mueen, A., and Keogh, E. (2016). Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM)*, pages 1317–1322. DOI: <https://doi.org/10.1109/ICDM.2016.0179>

Yu, S., Sun, X., Yu, Y., and Li, W. (2021). Frequencytemporal attention network for singing melody extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 251–255. DOI: <https://doi.org/10.1109/ICASSP39728.2021.9413444>

TO CITE THIS ARTICLE:

Plaja-Roglans, G., Nuttall, T., Pearson, L., Serra, X., & Miron, M. (2023). Repertoire-Specific Vocal Pitch Data Generation for Improved Melodic Analysis of Carnatic Music. *Transactions of the International Society for Music Information Retrieval*, 6(1), 13–26. DOI: <https://doi.org/10.5334/tismir.137>

Submitted: 05 April 2022 **Accepted:** 11 March 2023 **Published:** 26 June 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Transactions of the International Society for Music Information Retrieval is a peer-reviewed open access journal published by Ubiquity Press.