

Leveraging Melodic Context for Improved Svara Representation

Thomas Nuttall¹[0000–0001–6316–1424], Vivek Vijayan¹[0009–0002–9153–7868],
Xavier Serra¹[0000–0003–1395–2345], and Lara Pearson²[0000–0002–5073–8738]

¹ Music Technology Group, Universitat Barcelona Fabra, Barcelona, Spain

² Institute of Musicology, University of Cologne, Cologne, Germany

Corresponding author: `thomas.nuttall@upf.edu`

Abstract. For the South Indian musical tradition known as Carnatic music, embeddings of *svara* (note) pitch time series have proven useful for tasks such as svara classification and performance analysis. In this paper, we extend an existing embedding method by incorporating findings from musicological research on the relationship between the performance of a *svara* and its immediate melodic context, in order to improve the learning of these embedding models. We present a context-aware GRU-based model, adapting the existing DeepGRU architecture to encode both svara and its surrounding melodic context, before combining them via a co-attention mechanism prior to classification. For a ground truth dataset of 2,077 expert svara annotations across two performances in *rāga* Bhairavi, we observe that the inclusion of melodic context leads to a 6.6% absolute increase in F1 score for svara label classification (from 78.3% to 84.9%), and an 7.8% absolute increase (from 59.9% to 67.7%) for classification of *svara-form*: sub-svara clusters that capture *gamaka* (ornamentation) variations in the performed svara.

Keywords: Representation learning · Time series analysis · Carnatic music · Coarticulation

1 Introduction

Carnatic music, an art music tradition from South India, has been the subject of considerable recent research interest in Music Information Retrieval (MIR). Active areas of research on this musical style include melodic pattern recognition [14–16, 29, 30], note transcription [13, 40, 49, 52], music synthesis [39, 48] and performance analysis [17, 18, 31, 32]. These tasks often rely on latent representations, typically learnt by neural network models trained on expert-annotated data. Such representations offer a low-dimensional, length-invariant space that enables large-scale similarity comparisons across corpora.

However, note- or pattern-level annotated data are scarce, costly to obtain, and require input from multiple experts, whose annotations may diverge due to valid differences in level of detail. Transcription from audio to symbolic notation

is widely acknowledged in musicology as having a significant subjective component [11] with a typical issue in Carnatic music transcription being different possible approaches to the degree of detail represented in the notation [34].

A challenge in the automated estimation of melodic representations in this style is that svaras (note-level units) are often performed with gamakas (ornamentation), some of which oscillate around the theoretical pitch position without resting on it [22]. Furthermore, in any given raga, a svara may be performed with different gamakas, depending on its melodic context. As a result, there is a great deal of variability in the ways that a svara can be performed, making automated transcription particularly challenging in this style. However, the number of ways that a svara may be performed are not limitless, but rather are constrained by the tradition. Indeed, ragas are often best identified by their characteristic gamakas and motifs, rather than by their scalar content [53]. Therefore, there is scope for applying this knowledge that lies within the tradition to assist with svara identification.

In Carnatic music pedagogy, svaras and their associated gamakas are learnt in the context of musical phrases [53], and musicians themselves know that the way a svara is performed depends on its melodic context [48]. Following from these insights, the relationship between svara performance and immediate melodic context has been theorised using the concept of coarticulation: the tendency for the performance of a unit (in this case a svara) to be influenced by that which precedes or follows it [33]. Coarticulation is an important topic in phonetics, because phonemes differ in their realisation depending on their immediate context [24]. This indeed was one of the initial obstacles to automated speech transcription and synthesis [5, 25], a situation that can be aptly compared to that facing the computational analysis of Carnatic music.

The coarticulation hypothesis in Carnatic music has recently been investigated empirically using computational methods [31, 32]. This existing work creates a small dataset of svara annotations, as well as groupings of the different svara realisations into svara-gamaka units that are referred to as svara-forms. A first paper demonstrated a significant relationship between neighbouring svara and svara performance, and subsequent work found that the relationship increases as context increases, and plateaus after around 2 svaras either side.

In this paper, we present a novel methodology for improving svara representation learning by incorporating melodic context as an input to a svara embedding model. We demonstrate the value of this inclusion by evaluating the learned representations effectiveness for svara classification on an unseen ground-truth test set from the original annotations.

The code to reproduce this analysis and access the Context-Aware DeepGRU model can be found in the accompanying Github repository³.

³ <https://github.com/thomasgnuttall/ContextGRU/>

2 Related Work

Latent representations of sequential data have proven to be valuable intermediates for various music analysis tasks, including transcription [13], motif finding [30, 56], classification [23, 26, 37, 44, 46], performance analysis [31, 32], and synthesis [38], both in Carnatic music and beyond. In many cases, the learned representation serves as a distilled encoding useful across multiple tasks [2, 51, 55].

A common limitation in training these models is the lack of annotated ground truth data, and whilst some small datasets have been published [17, 30, 32], in Carnatic music, this remains a problem. A comprehensive review of deep learning techniques for time series with limited labels can be found in [10], and can be broadly grouped into three categories: developing model components with fewer parameters to improve performance on small datasets, such as the case with Gated Recurrent Units (GRUs) (shown to perform well in classifying *svara* with limited data) [27, 31]; by pretraining in a self-supervised fashion on a broader, related dataset [6, 21, 46]; or by leveraging more information from existing data, such as including additional contextual features [8] or modalities [7, 54] to contextualize the primary input. It is the latter that this paper is concerned with.

Due to the ornamented nature of the style, Carnatic *svaras* are well characterized by segments of continuous pitch time series [1, 17–20, 22, 29, 35, 36, 41, 42, 53]. Deep learning models for time series representation do not typically incorporate explicitly defined contextual features [9, 27, 28]. However, for the case of pitch time series of *svara*, recent musicological and computational studies have shown that neighbouring melodic context provides important information about how individual *svaras* are performed [31–33].

In this paper we leverage the findings in [31–33] to improve *svara* representation. We do this by adapting an existing embeddings model [27, 31] to incorporate contextual information provided by the time series surrounding ground truth *svara* annotations, demonstrating its effectiveness for the task of *svara* label classification.

3 Dataset

We work with two performances in *rāga* Bhairavi from the Carnatic corpus of the Saraga dataset: Kamakshi (composed by Syama Sastri), performed by Sanjay Subrahmanyam, and Raksha Bettare (composed by Tyagaraja), performed by Shruthi S. Bhat [43, 47]. The total duration of these composition performances is approximately 25 minutes and they are accompanied by a total of 2,077 expert *svara* annotations, created collaboratively by two Carnatic musicians. The annotations are made in sargam notation, a form of notation traditionally used in Carnatic music that is similar to solfège (the seven *svaras* in this *rāga* are sa, ri, ga, ma, pa, dha, and ni). Of these 2,077 *svara* annotations, 804 *svaras* include an additional *svara-form* label, denoting sub-*svara* clusters that capture gamaka variations, such as the four ga variants in Fig. 1. All annotations were

manually verified by a Carnatic music expert and have proven useful for music analysis [31,32].

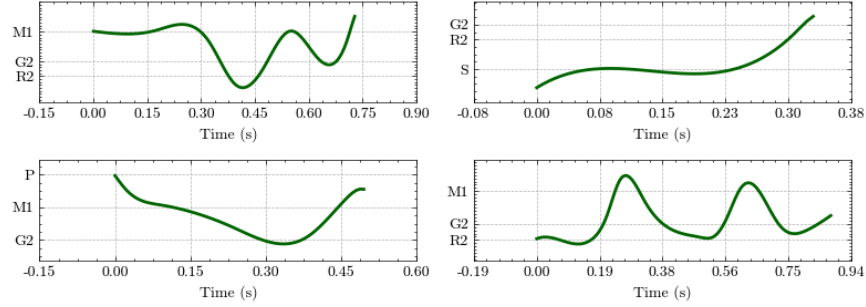


Fig. 1. Predominant pitch curves for four variants of the svara ga in *rāga* Bhairavi. These plots illustrate how the same svara can be performed with a number of different gamakas - melodic movements - that traverse a range of pitches.

4 Methodology

In this section we detail the extraction of the f0 time series corresponding to the predominant vocal melody for each of the ground truth annotations and their surrounding context, and present a GRU-based recurrent neural network for encoding and classifying svara alongside its context.

4.1 Svara Time Series

Each annotation in our dataset corresponds to a performed svara in one of two compositions in *rāga* Bhairavi. For both performances, we extract the f0 time series corresponding to the predominant vocal melody using the FTA-Net Carnatic pitch extraction model available in the compIAM package [12]. Short silences (up to 200 ms) are interpolated, and pitch is expressed in cents relative to the performer’s tonic, as provided by the Saraga dataset. We smooth the pitch curves using a cubic spline, preserving melodic peaks.

Our classification model expects individual svara observations, which it encodes and classifies (Section 4.2). Each observation comprises three distinct time series: (1) the f0 pitch curve of the *current* svara annotation, (2) the f0 pitch curve of the preceding melodic context, and (3) the f0 pitch curve of the succeeding melodic context (see the upper section of Fig. 2). Current svaras containing intermediate silence (i.e., not at the boundaries) are excluded due to the likelihood that they have pitch extraction errors.

The two contextual pitch curves are extracted from segments preceding and succeeding the current svara, with a duration twice that of the svara itself. This

follows [32], which shows diminishing correlation between svara performance and context beyond two adjacent svaras. If the contextual segments exceed one second, they are capped at one second. If silence is present, contextual segments are trimmed to start or end with silence for preceding and succeeding contexts, respectively.

Whilst silence in the pitch curve of the central, *current*, annotation is likely due to pitch extraction errors and thus excluded, silence in the preceding and succeeding context is expected (e.g., ends of phrases) and treated as being meaningful (potentially having its own influence). We retain this silence by substituting it with a pitch value far outside the expected pitch range of a Carnatic vocalist (-4000 cents) and by introducing a binary feature indicating silence. Consequently, each of the three pitch curve time series per observation is two-dimensional: one feature representing pitch (in cents), and one binary feature indicating silence.

Each observation is associated with two labels: the svara label, indicating the performed svara ($\in \{\text{sa, ri, ga, ma, pa, dha, ni}\}$), and the svara-form label, representing the cluster of unique gamaka (Section 3) [32]. Fig. 1 illustrates four observations for the svara ga.

4.2 Context-Aware DeepGRU Network

Our model (Fig. 2) is an adaptation of the existing DeepGRU model [27], which has been demonstrated as effective in classifying svara in small datasets [31]. To allow the model to focus on temporally salient information within and beyond the primary sequence of interest, we alter the attention mechanism to apply it independently to the outputs of three GRU-based encoders: *preceding*, *current*, and *succeeding*. Consisting of stacked GRUs, each encoder processes the variable-length time series pitch curves corresponding to the preceding context, current svara, and succeeding context respectively (Section 4.1), to produce a sequence of hidden representations. Attention is computed on these representations with reference to the final hidden state of the *current* encoder.

If $H \in \mathbb{R}^{B \times T \times d}$ denotes the sequence of hidden states produced by an encoder - where B is the batch size, T the sequence length, and d the hidden dimensionality - and $h_c \in \mathbb{R}^{B \times d}$ represents the final hidden state of the current encoder, the attention mechanism computes a compatibility score between each time step in H and the reference state h_c , projected via a learnable linear transformation $W \in \mathbb{R}^{d \times d}$:

$$e_t = H_t W h_c^\top, \quad \text{for } t = 1, \dots, T$$

These scores are normalized via a softmax function over the temporal dimension to obtain attention weights $\alpha \in \mathbb{R}^{B \times T \times 1}$:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)}$$

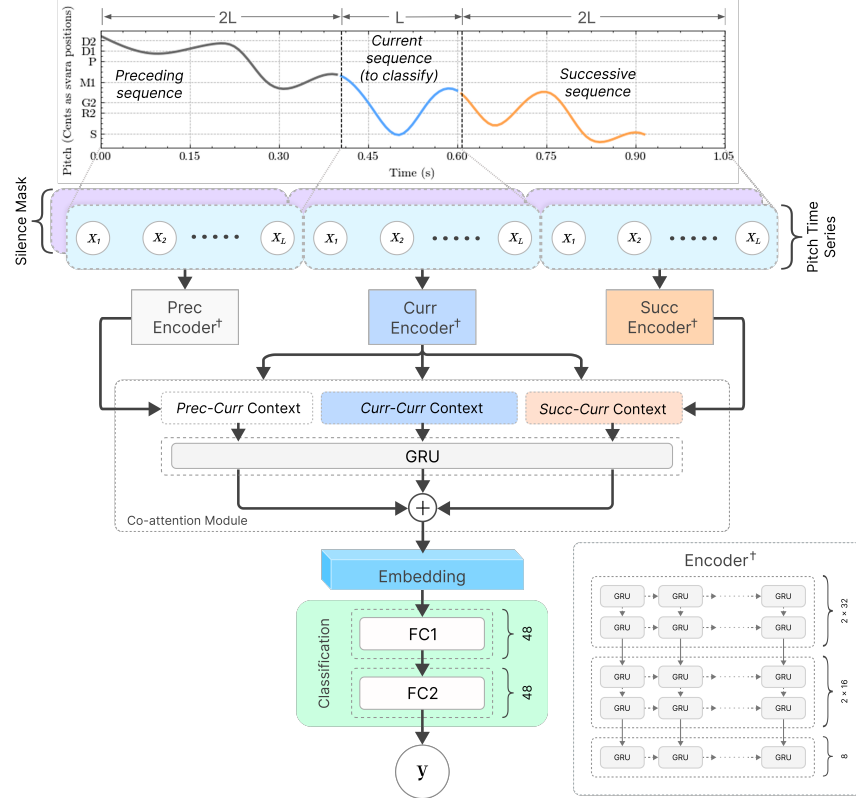


Fig. 2. Context-aware DeepGRU Network. Contextual and current pitch time series are encoded separately and combined through a co-attention mechanism before classification.

And the context vector $c \in \mathbb{R}^{B \times d}$ is computed as the weighted sum of the input sequence:

$$c = \sum_{t=1}^T \alpha_t H_t$$

As in the original DeepGRU model, to capture higher-order temporal interactions, this context vector is passed to an auxiliary single-layer GRU, initialized with h_c , yielding a gated context representation c' . The final attention output is then formed by concatenating the raw context vector and the gated context:

$$\tilde{c} = [c \parallel c']$$

This process is applied independently to the preceding, current, and succeeding encoders' outputs, with all attention computations referencing the current encoder's final hidden state, h_c . This ensures that context integration remains dynamically aligned with the current sequence, and enables the model to selectively attend to temporally relevant cues from the two contextual sequences.

The classification head of the original DeepGRU model remains unchanged - the three attention outputs are concatenated and passed to two fully connected layers with batch normalization, dropout, and ReLU activation functions, mapping the learned feature representations to the final output classes.

5 Experiments and Results

5.1 Experiment 1: Ablation study

To understand the impact of melodic context on *svara* classification, we train two models on the ground truth annotations: one with contextual data, using the model presented in Section 4.2, and one using only the central *current* time series, with the contextual encoders removed. In the latter case, the model reduces to the original DeepGRU architecture. Each model is trained twice—once to predict *svara*-form and once to predict *svara*—resulting in four experiments in total.

Each encoder is configured with an input layer size of 32, halving in size at each subsequent layer. We use the Adam optimizer, a learning rate of 10^{-3} , and a weight decay of 10^{-4} . All experiments use a mini-batch size of 256 and a dropout rate of 0.3. These training parameters follow the original DeepGRU PyTorch implementation. Models are evaluated using 3-fold cross-validation, and we report the average F1 score on the test set across all folds. To prevent overfitting, training data is augmented by balancing class distributions using time dilation with a factor of between 0.9 and 1.1. Table 1 displays the results.

We note that for predicting both *svara*-form and *svara*, the inclusion of melodic context significantly improves model performance — a **6.6%** absolute increase in F1 score for *svara* classification (from 78.3% (± 2.6) to 84.9% (± 2.5)), and a **7.8%** absolute increase (from 59.9% (± 0.6) to 67.7% (± 3.3)) for *svara*-form.

Table 1. Effect of considering melodic context on Svara and Svara-form Classification

Melodic Context	Target	F1 mean	F1 std
No	Svara	0.783	0.026
Yes	Svara	0.849	0.025
No	Svara-form	0.599	0.006
Yes	Svara-form	0.677	0.033

5.2 Experiment 2: Encoder Contributions

To understand the relative impact of the preceding, succeeding, and current encoder on the final prediction, we conduct a gradient-based attribution analysis to measure how sensitive the model’s prediction is to small perturbations in specific encoder outputs, quantified via the L2 norm of the gradient. Specifically, we compute the gradient of the predicted class score with respect to the attention output of each encoder: preceding context, current input, and succeeding context. The L2 norm of each gradient is used as a proxy for the encoder’s influence on the model’s decision, with higher norms indicating greater sensitivity. This approach has been demonstrated to be useful for comparing internal components’ influence without retraining or ablation [3, 4, 45, 50].

Averaged over the entirety of the ground truth data, the normalized gradient norms are 0.295, 0.402, 0.303 for the preceding, current, and succeeding encoder respectively, suggesting that whilst each encoder contributes meaningfully, the central input encoder has the highest influence on the classification decision, with the two contextual encoders demonstrating a smaller but equally balanced contribution. It is perhaps not surprising that the central input should dominate, with contextual information providing supportive but secondary contributions.

6 Conclusion

In this paper we present a context-aware, GRU-based time series classifier, adapting the existing DeepGRU to encode multiple time series separately and combine them via a co-attention mechanism before classification. We show the effectiveness of this model in contextualizing svara embeddings with their surrounding melodic information by improving the F1 score of svara classification by 6.6%, (from 78.3% (± 2.6) to 84.9% (± 2.5)), and 7.8% (from 59.9% (± 0.6) to 67.7% (± 3.3)) for svara-form, on a ground truth dataset of 2,077 expert annotations.

This finding also corroborates previous musicological and computational studies on the relationship between svara performance and immediate melodic context, and explicates a characteristic of the style implicitly understood by practitioners.

Given the widespread use of representation learning for MIR tasks, we anticipate that improvements to these models, such as that presented in this paper, will positively impact downstream applications such as transcription, pattern recognition, and synthesis, by more effectively capturing this implicit musical knowledge embedded within the Carnatic tradition.

Acknowledgments. This research was carried out as part of the "IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA.

References

1. Akant, K.: Measuring frequencies of shrutis in indian classical music. In: 2019 9th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-19). pp. 1–4. IEEE (2019)
2. Alonso-Jiménez, P., Bogdanov, D., Pons, J., Serra, X.: Tensorflow audio models in Essentia. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)
3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: International Conference on Learning Representations (ICLR) (2018)
4. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K.R., Müller, K.R.: How to explain individual classification decisions. *Journal of Machine Learning Research* **11**, 1803–1831 (2010)
5. Birkholz, P.: Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLOS ONE* **8**(4), e60603 (2013). <https://doi.org/10.1371/journal.pone.0060603>
6. Bošnjak, M., Richemond, P.H., Tomasev, N., Strub, F., Walker, J.C., Hill, F., Buesing, L.H., Pascanu, R., Blundell, C., Mitrovic, J.: Semppl: Predicting pseudo-labels for better contrastive representations. *arXiv preprint arXiv:2301.05158* (2023)
7. Clayton, M., Rao, P., Shikarpur, N.N., Roychowdhury, S., Li, J.: Raga classification from vocal performances using multimodal analysis. In: ISMIR. pp. 283–290 (2022)
8. Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* (2024)
9. Ekambaram, V., Kumar, S., Jati, A., Mukherjee, S., Sakai, T., Dayama, P., Gifford, W.M., Kalagnanam, J.: Tspulse: Dual space tiny pre-trained models for rapid time-series analysis. *arXiv preprint arXiv:2505.13033* (2025)
10. Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwok, C.K., Li, X.: Label-efficient time series representation learning: A review. *IEEE Transactions on Artificial Intelligence* **5**(12), 6027–6042 (2024). <https://doi.org/10.1109/TAI.2024.3430236>
11. Ellingson, T.: Transcription. In: Myers, H. (ed.) *Ethnomusicology: an introduction*, pp. 110–52. *Ethnomusicology: an introduction*, Norton, New York (1992)
12. Genís Plaja-Roglans and Thomas Nuttall and Xavier Serra: compiam (2023), <https://mtg.github.io/compIAM/>
13. Gowrishankar, B., Bhajantri, N.U.: Deep learning long short-term memory based automatic music transcription system for carnatic music. In: 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). pp. 1–6. IEEE (2022)
14. Gulati, S.: Computational approaches for melodic description in indian art music corpora. Ph.D. thesis, Universitat Pompeu Fabra (2016)

15. Gulati, S., Serra, J., Ishwar, V., Serra, X.: Mining melodic patterns in large audio collections of indian art music. In: 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems. pp. 264–271. IEEE (2014)
16. Ishwar, V., Dutta, S., Bellur, A., Murthy, H.A.: Motif spotting in an alapana in carnatic music. In: ISMIR. pp. 499–504 (2013)
17. Koduri, G.K., Ishwar, V., Serrà, J., Serra, X.: Intonation analysis of rāgas in carnatic music. *Journal of New Music Research* **43**(1), 72–93 (2014)
18. Koduri, G.K., Serrà Julià, J., Serra, X.: Characterization of intonation in carnatic music by parametrizing pitch histograms (2012)
19. Komaragiri, M.M.: An empirical study of intonation in raga kalyani (2011)
20. Komaragiri, M.M.: Pitch analysis in South Indian music: with a critical examination of the theory of 22 śruti-s (2013)
21. Krause, M., Weiß, C., Müller, M.: A cross-version approach to audio representation learning for orchestral music. In: ISMIR. pp. 832–839 (2023)
22. Krishna, T.M., Ishwar, V.: Carnatic music: Svara, gamaka, motif and raga identity. In: Serra, X., Rao, P., Murthy, H., Bozkurt, B. (eds.) *Proceedings of the 2nd CompMusic Workshop*, pp. 12–18. Universitat Pompeu Fabra, Barcelona (2012)
23. Krishnendu, R., et al.: Classification of carnatic music ragas using rnn deep learning models. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2023)
24. Kühnert, B., Nolan, F.: The origin of coarticulation. In: Hardcastle, W.J., Hewlett, N. (eds.) *Coarticulation*, pp. 7–30. Cambridge University Press, 1 edn. (Dec 1999). <https://doi.org/10.1017/CBO9780511486395.002>
25. Lee, K.F.: Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition **38**(4), 599–609
26. Madhusudhan, S.T., Chowdhary, G.: Deepsrgm-sequence classification and ranking in indian classical music with deep learning. arXiv preprint arXiv:2402.10168 (2024)
27. Maghoumi, M., Jr., J.J.L.: Deepgru: Deep gesture recognition utility. *CoRR abs/1810.12514* (2018), <http://arxiv.org/abs/1810.12514>
28. Malhotra, P., TV, V., Vig, L., Agarwal, P., Shroff, G.: Timenet: Pre-trained deep recurrent neural network for time series classification. arXiv preprint arXiv:1706.08838 (2017)
29. Nuttall, T., Plaja-Roglans, G., Pearson, L., Serra, X.: The matrix profile for motif discovery in audio - an example application in carnatic music. In *Proc. of the 15th Int. Symposium on Computer Music Multidisciplinary Research (CMMR)*, Tokyo, Japan pp. 109–118 (2022)
30. Nuttall, T., Plaja-Roglans, G., Pearson, L., Serra, X.: In search of sañcāras: Tradition-informed repeated melodic pattern recognition in carnatic music. In: *In Proceedings of the 23rd International Conference on Music Information Retrieval (ISMIR)*, Bengaluru, India. pp. 337–344. <https://repositori.upf.edu/handle/10230/56440>
31. Nuttall, T., Serra, X., Pearson, L.: Svara-forms and coarticulation in carnatic music: an investigation using deep clustering. In: *Proceedings of the 11th International Conference on Digital Libraries for Musicology*. pp. 15–22 (2024)
32. Nuttall, T., Serra, X., Pearson, L.: Svara-forms in carnatic music: Contextual influences on the performance of svara. In: *Workshop on Indian Music Analysis and Generative Applications (WIMAGA)*, a Satellite Workshop of the IEEE International Conference on Acoustics, Speech, and Signal Processing (2025)

33. Pearson, L.: Coarticulation and gesture: an analysis of melodic movement in South Indian raga performance. *Music Analysis* **35**(3), 280–313 (2016), <https://doi.org/10.1111/musa.12071>
34. Pearson, L., Manickavasakan, B.: Annotating Karnataka Music: Encounters Between a Musical Tradition and Computational Tools. In: Bonini Baraldi, F. (ed.) *Second Symposium of the ICTM Study Group on Sound, Movement, and the Sciences (SoMoS)*. pp. 23–27. Barcelona, Spain (2023), <https://zenodo.org/records/10423805>
35. Pearson, L., Nuttall, T., Pouw, W.: Landscapes of coarticulation: The co-structuring of gesture-vocal dynamics in karnatak vocal performance (2024), <https://osf.io/npm96>
36. Plaja-Roglans, G., Nuttall, T., Pearson, L., Serra, X., Miron, M.: 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 (2023)
37. Rajan, R., Sivan, S.: Multi-channel cnn-based rāga recognition in carnatic music using sequential aggregation strategy. *Circuits, Systems, and Signal Processing* **42**(7), 4072–4095 (2023)
38. Sankaranarayanan, R., Heck, L., Weinberg, G.: Gamaka synthesis for kalpitha swaras in carnatic music. In: *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. pp. 1–5. IEEE (2025)
39. Sankaranarayanan, R., Weinberg, G.: Gamaka Synthesis for Kalpitha Swaras in Carnatic Music. In: *Extended Abstracts for the Late-Breaking Demo Session of the 23rd International Society for Music Informaion Retrieval Conference Bengaluru, India, 2022*. Bengaluru, India (2022)
40. Sekhar, P.K., Viraraghavan, V.S., Sankaran, S., Murthy, H.A.: An approach to transcription of varnams in carnatic music using hidden markov models. In: *2017 Twenty-third National Conference on Communications (NCC)*. pp. 1–6. IEEE (2017)
41. Sentürk, S., Koduri, G.K., Serra, X.: A score-informed computational description of svaras using a statistical model. In: *13th Sound and Music Computing Conference, Hamburg, Germany* (2016)
42. Serra, J., Koduri, G.K., Miron, M., Serra, X.: Assessing the tuning of sung indian classical music. In: *ISMIR*. pp. 157–162. Florida (2011)
43. Serra, X.: Saraga audiovisual: a large multimodal open data collection for the analysis of carnatic music (2024)
44. Shah, D.P., Jagtap, N.M., Talekar, P.T., Gawande, K.: Raga recognition in indian classical music using deep learning. In: *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*. pp. 248–263. Springer (2021)
45. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2014)
46. Spijkervet, J., Burgoyne, J.A.: Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410* (2021)
47. Srinivasamurthy, A., Gulati, S., Repetto, R.C., Serra, X.: Saraga: Open datasets for research on indian art music. *Empirical Musicology Review* **16**(1), 85–98 (2021)
48. Subramanian, M.: Carnatic music - automatic computer synthesis of gamakams. *Journal of the Sangeet Natak Akademi* **XLIII**(3), 28–36 (2009)

49. Suma, S., Koolagudi, S.G., Ramteke, P.B., Rao, K.: Note transcription from car-natic music. In: Smart Computing Paradigms: New Progresses and Challenges: Proceedings of ICACNI 2018, Volume 1. pp. 123–129. Springer (2020)
50. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (ICML). pp. 3319–3328. PMLR (2017)
51. Talukder, S.J., Yue, Y., Gkioxari, G.: TOTEM: TOkenized time series EMbeddings for general time series analysis. Transactions on Machine Learning Research (2024), <https://openreview.net/forum?id=QlTLkH6xRC>
52. Viraraghavan, V.S., Pal, A., Murthy, H., Aravind, R.: State-based transcription of components of car-natic music. In: ICASSP 2020-2020 IEEE International Confer-ence on Acoustics, Speech and Signal Processing (ICASSP). pp. 811–815. IEEE (2020)
53. Viswanathan, T.: The analysis of rāga ālāpana in south indian music. Asian Music pp. 13–71 (1977)
54. Wadhwa, L., Mukherjee, P.: Music genre classification using multi-modal deep learning based fusion. In: 2021 Grace Hopper Celebration India (GHCI). pp. 1–5. IEEE (2021)
55. Wang, Z., Xia, G.: Musebert: Pre-training music representation for music under-standing and controllable generation. In: ISMIR. pp. 722–729 (2021)
56. Wu, Y., Dannenberg, R.B., Xia, G.: Motif-centric representation learning for sym-bolic music. arXiv preprint arXiv:2309.10597 (2023)