Uncovering Carnatic Vocal Gestures via Pitch-Driven Note Segmentation: A Leave-One-Singer-Out Analysis

Anish Nair, Aaron Lanterman

April 2025

Abstract

We present a multimodal method to predict high versus low-pitch notes in Carnatic vocals solely from visual gestures. First, individual notes are segmented from continuous audio using Librosa's `pyin` pitch contour and onset detection. Each note is then synchronized with upper-body and facial features extracted via Mediapipe pose (hand height, torso lean, elbow angle, head tilt) and FaceMesh (jaw opening, eye opening, mouth width). We evaluate five standard classifiers: logistic regression, random forest, SVM, gradient boosting, and MLP in a leave-one-singer-out (LOOV) protocol across five professional vocalists. Our results demonstrate strong singer independent generalization and clear improvements from simple hyperparameter tuning. We introduce a feature importance stability metric (cosine distance between training and held-out importance vectors) and show that singers whose importance patterns remain more consistent across splits achieve markedly better generalization. Finally, by reducing FRAME_SKIP to 1 and limiting tree depths to 3, RF AUC rises to 0.66 $\pm$ 0.04 and logistic to 0.71 $\pm$ 0.03. Across all experiments, head tilt and hand posture emerge as the most predictive visual cues.

1 Introduction

Music and Gesture research has long established that performers' bodily movements during a musical performance convey expressive and structural information beyond the acoustical signal [4, 5]. In western classical and popular music, studies have shown that hand gestures, head nods, and torso sways correlate with phrasing, emphasis, and entrainment [3, 12]. However, comparatively little work has explored these phenomena in South Asian classical traditions. Carnatic music is a highly ornamented and improvisatory art music tradition from South India. Vocalists navigate complex rāga (melodic mode) and tāḷa (rhythmic cycle) structures, often without bar lines or strict time signatures, and rely on subtle pitch inflections (gamaka), microtonal ornamentation, and long melodic phrases to convey meaning and emotion. Ethnographic studies [10, 7]

show that carnatic vocalists routinely use hand-raising, finger pointing, and head movements to convey rāga phrases, cue accompanists, and signal structural divisions (e.g. muktāyi swara vs. niraval passages). Unlike Western classical conducting gestures, these singer-generated movements are idiosyncratic and tied directly to melodic motion. In Western Music Information Retrieval (MIR) Literature, skeleton-based action recognition and pose-driven gesture analysis have been applied to orchestral conducting [6] and pop dance stimuli [12] and recently the MIR community has begun to leverage pose estimation for Indian classical vocals, most notably in a Hindustani raga and singer classification study that used 2D/3D skeleton data to distinguish three performers [2], but there are no comparable computational investigations focused on Carnatic vocals. Recent advances in pose-estimation using computer vision algorithms (e.g. Mediapipe, OpenPose) [8, 1] and audio analysis (Librosa's pyin) [9] make it possible to extract high-dimensional gesture and pitch features from ordinary video and audio recordings. In this paper, we introduce a novel pitch-driven note-segmentation pipeline for Carnatic vocal performances that aligns Librosa's pyin-derived f_0 contours with upper-body and facial gestural features extracted via Mediapipe's Pose (hand height, torso lean, elbow angle, head tilt) and FaceMesh (jaw opening, eye opening, mouth width) modules. In addition, we evaluate the singer-independence of these visual cues in a leave-one-singer-out framework—training and testing five classifiers (logistic regression, random forest, SVM, gradient boosting, MLP) on five professional Carnatic vocalists—and analyze feature-importance stability across performers.

2 Dataset

For our visual–acoustic gesture study, we utilize the **Sanidha** dataset [13], a studio quality, multimodal corpus of 5 different Carnatic performances. Each performance includes professionally recorded, source-separated audio tracks for vocals, mridangam, ghatam, and violin, all synchronized with high-definition video. The **key properties** of the dataset in relation to this study includes:

- **Five professional vocalists:**
  - Two female: Anjana Nagarajan, Amita Krishnan
  - Three male: Prasanna Soundararajan, Prashant Krishnamoorthy, Salem Shriram

  We hold out *Salem Shriram* for validation in our LOOV experiments.
- **Audio–video recordings:** Each singer's performance was captured in a modern isolation studio with
  - CD-quality audio (44.1 kHz, 16 bit), delivered as *ground-truth* isolated stems (vocals and accompaniment) for source-separation tasks.
  - Synchronized 1080p@29.97 fps video, front-view of the singer's upper body and face.

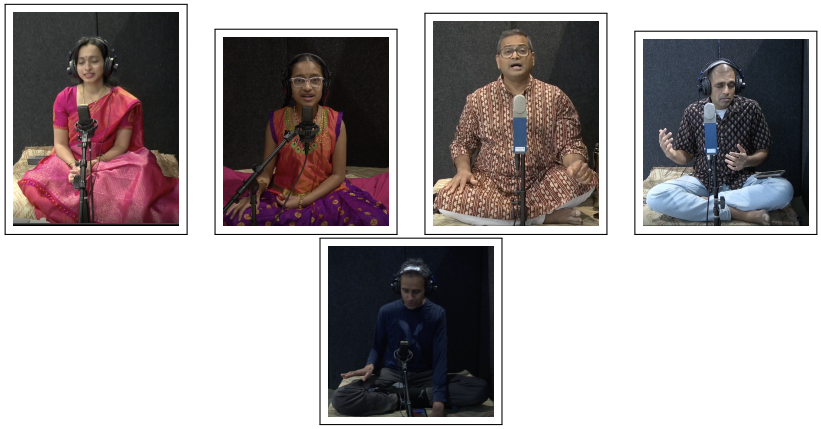

Figure 1: Front-view snapshots of our five singers (Anjana, Amita, Prasanna, Prashant, Salem).

- **3-minute excerpts:** From each vocalist we extract contiguous 3 min segments containing representative improvisational material (alapana, neraval, kalpana swaram).
- **Diversity of performers:**
  - *Gender balance:* two female, three male vocalists.
  - *Age & experience:* artists range from early-career (10-20), to highly experienced performers (40s–50s), which ensures a broad diversity of performers with different stylistic and training backgrounds
  - *Melodic–metric variety:* the excerpts cover a wide selection of rāgas (e.g. Kalyāni, Bhairavi, Latanga, Mohanam) and tālas (Adi, Rupaka, Misra-Cāpu).
- **Ground-truth separation:** Unlike prior live-concert corpora (e.g. Saraga), Sanidha was recorded with each musician in an acoustically isolated room. This gives us *clean multi-track* stems with negligible bleed.

3 Experiment

3.1 Feature Selection

We chose four upper body pose features: hand height, torso lean, elbow angle, and head tilt and three FaceMesh features: jaw opening, eye opening, mouth width based on two criteria:

1. **Empirical coupling to pitch:** Motion-capture analyses in Carnatic vocal performance (Pearson & Pouw, 2022) demonstrated that these specific

kinematic variables exhibit the strongest temporal and magnitude coupling to F_0 changes, and proving the most relevant predictors of pitch inflection.

2. **Reliable video extraction:** All seven landmarks can be robustly and efficiently extracted from ordinary 2D concert footage using Mediapipe's Pose and FaceMesh pipelines without sensor rigs.

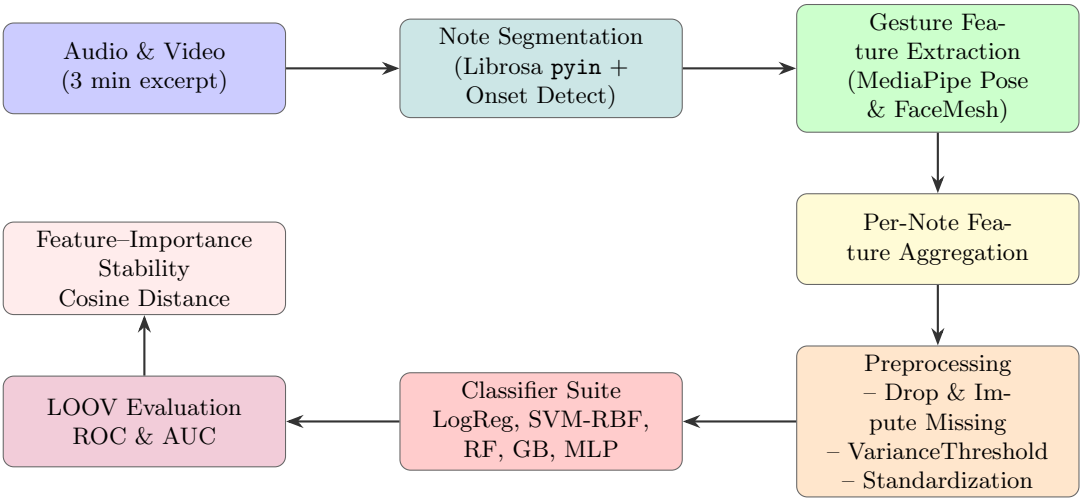

Figure 2: Complete model architecture for Carnatic gesture–pitch analysis

3.2 Model Architecture

Let $y(t)$ be the continuous audio signal sampled at rate f_s, and let $\{F_i\}_{i=1}^{T}$ be the corresponding video frames at times t_i. Our goal is to map each short note interval to a binary *high/low* pitch label via visual features alone. The pipeline proceeds in five stages:

(1) Note segmentation. We first extract a pitch-contour $F_0(n)$ via Librosa's `pyin`:

$$F_0(n) \;=\; \mathrm{pyin}\big(y[\,n\,]\big), \quad n = 1, \ldots, N,$$

and detect frame-level onset indices $\mathcal{O} = \{o_k\}$ by onset_detect(y). These define note intervals $\big[\tau_k, \tau_{k+1}\big)$ with $\tau_k = o_k / f_s$.

(2) Per-frame gesture features. For each video frame at time t, let $\mathbf{p}(t) \in \mathbb{R}^{33}$ be the 3D pose landmarks from MediaPipe Pose, and $\mathbf{f}(t) \in \mathbb{R}^{468}$ the

FaceMesh landmarks. We compute:

$$h(t) = y_{\text{shoulder}}(t) - y_{\text{wrist}}(t), \quad \ell(t) = y_{\text{mid shoulders}}(t) - y_{\text{mid hips}}(t),$$
$$\theta(t) = \cos^{-1}\Big(\tfrac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\Big), \quad \phi(t) = y_{\text{shoulder mid}}(t) - y_{\text{nose}}(t),$$
$$j(t) = \|\mathbf{f}_{13}(t) - \mathbf{f}_{14}(t)\|, \quad e(t) = \|\mathbf{f}_{159}(t) - \mathbf{f}_{145}(t)\|, \quad m(t) = \|\mathbf{f}_{61}(t) - \mathbf{f}_{291}(t)\|.$$

Stacking these into $\mathbf{x}(t) = [\, h, \ell, \theta, \phi, j, e, m \,]^\top \in \mathbb{R}^7$.

(3) Note-level aggregation. For each note interval $\mathcal{I}_k = \{t : \tau_k \le t < \tau_{k+1}\}$, we form a 7-dimensional summary

$$\mathbf{z}_k \;=\; \frac{1}{|\mathcal{I}_k|} \sum_{t \in \mathcal{I}_k} \mathbf{x}(t).$$

We also record its mean pitch $\bar{F}_0^{(k)}$ and assign a binary label $y_k = \mathbf{1}\{\bar{F}_0^{(k)} > \text{median}(\bar{F}_0)\}$.

(4) Preprocessing. Collect all M note vectors into $X \in \mathbb{R}^{M \times 7}$.

1. *Feature drop:* remove column j if more than a fraction α of its entries are missing.
2. *Imputation:* fill remaining NaNs by column medians.
3. *Variance threshold:* keep only those columns with $\text{Var}(X_{:,j}) > \epsilon$.
4. *Standardization:* for each column j, $\tilde{X}_{ij} = (X_{ij} - \mu_j)/\sigma_j$.

(5) Classification & LOOV evaluation. We compare five learners:

$$\begin{aligned}
&\text{Logistic regression:} && p(y = 1 \mid \mathbf{z}) = \sigma(\mathbf{w}^\top \mathbf{z} + b),\\
&\text{SVM (RBF):} && \min_{\alpha} \; \tfrac{1}{2}\alpha^\top Q \alpha - \mathbf{1}^\top \alpha,\\
&\text{Random forest:} && \hat{f}(\mathbf{z}) = \tfrac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{z}),\\
&\text{Gradient boosting:} && F_M(\mathbf{z}) = \sum_{m=1}^{M} \gamma_m \, h_m(\mathbf{z}),\\
&\text{MLP:} && f(\mathbf{z}) = W^{(2)} \sigma\big(W^{(1)} \mathbf{z} + b^{(1)}\big) + b^{(2)}.
\end{aligned}$$

In a leave-one-singer-out (LOOV) protocol we iteratively hold out all notes from singer s, train on the rest, and compute ROC/AUC on the held-out set. This directly measures *singer-independent* generalization.

(6) Feature-importance stability. For random forest we extract normalized Gini importances $\mathbf{i}_{\text{train}}$ and $\mathbf{i}_{\text{holdout}}$. We then compute their cosine distance

$$D = 1 - \frac{\mathbf{i}_{\text{train}} \cdot \mathbf{i}_{\text{holdout}}}{\|\mathbf{i}_{\text{train}}\| \, \|\mathbf{i}_{\text{holdout}}\|},$$

which provides a scalar measure of how stable each feature's ranking is across singers. The ROC and the feature importance with each singer as a "hold-out" is analyzed below:

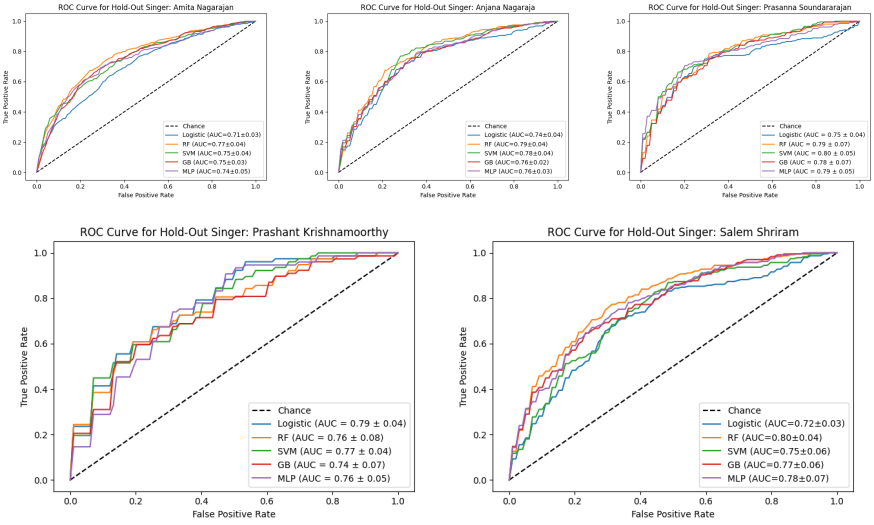

Figure 3: ROC curves for each singer held out in the LOOV evaluation. Each panel shows TPR vs. FPR for logistic regression, SVM, RF, GB, and MLP.

Across all hold out singers, Random Forest (RF) edges out the other classifiers on four splits, with Logistic Regression taking the lead on the remaining one, yet the AUC gaps are small (only 0.02–0.03).

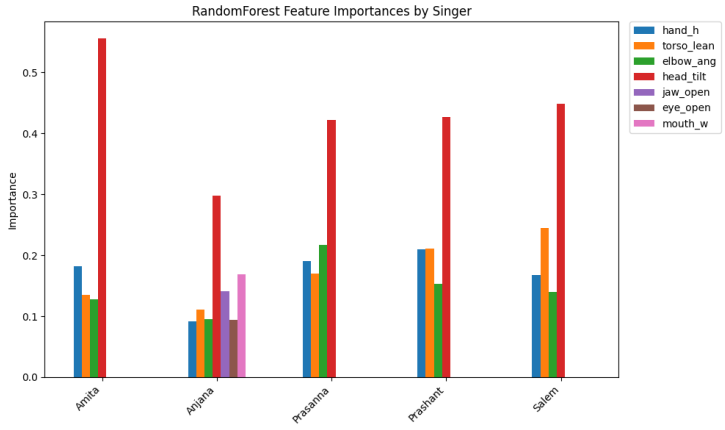

Figure 4: Random Forest feature importances for each singer in the LOOV experiment. Features are ranked by mean decrease in Gini impurity.

Because Random Forest led in four of five LOOV folds, we examined its Gini importance scores to see which gestures drive its predictions. As Figure 4 shows, head tilt carries the most weight. Vertical head motion most closely mirrors pitch contours where ascending notes coincide with upward tilts and descending notes with downward tilts, while hand elevation ranks a close second. The same would go for hand movement. Elevation of hand would increase pitch and decrease vis-a-versa.

3.3 Hyperparameter Tuning and Validation

Building on our Baseline ROC and Feature Importance curves which includes the TPF and FPR for the ground truth validation singer, we have created a LOOV pipeline to train on the 4 singers and predict high and low pitch notes of our validation singer based purely on their gesture.

Baseline settings.

- **FRAME_SKIP**: 2 (process every other video frame)
- **Random Forest (RF) max_depth**: 7
- **Gradient Boosting (GB) max_depth**: 5

Under our original configuration, we processed every other video frame and allowed trees up to depth 7 (RF) and 5 (GB). Figure 5 shows the resulting LOOV AUCs for each classifier on the held-out singer Salem:

Logistic: 0.69±0.05, RF: 0.63±0.04, SVM: 0.61±0.06, GB: 0.60±0.05, MLP: 0.67±0.06.

Figure 6 plots RF AUC against feature-importance cosine distance for each

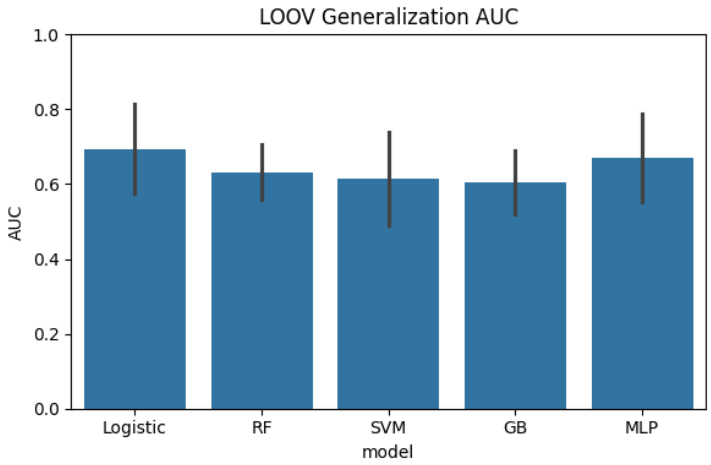

Figure 5: LOOV generalization AUC under baseline hyperparameters.

singer, confirming that those with more stable importance rankings (lower distance) generalize better—Salem sits at ($D = 0.10,\ \text{AUC} = 0.72$).

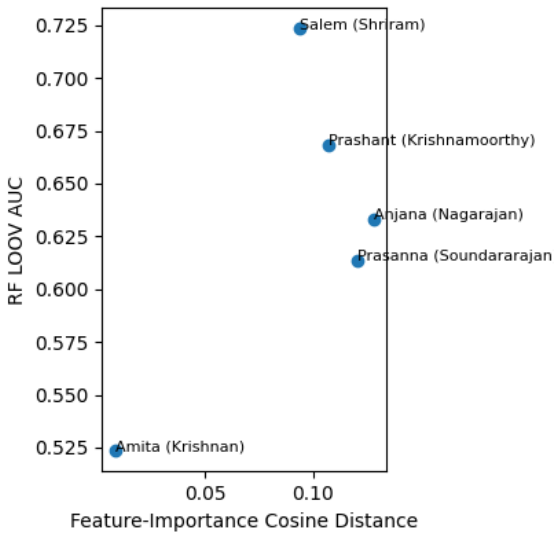

Figure 6: RF AUC vs. feature-importance cosine distance under baseline settings.

Finally, the confusion matrix for RF on Salem (Fig. 7) yields 135 true negatives, 2 false positives, 115 false negatives, and 22 true positives

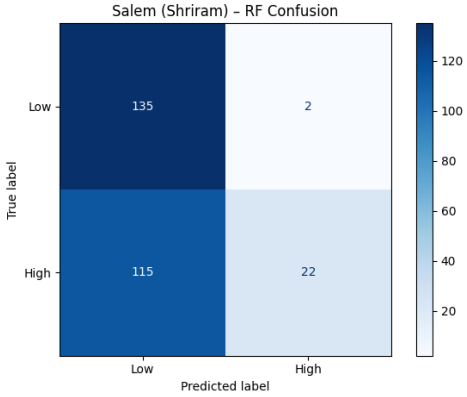

Figure 7: Confusion Matrix of Predicted Label vs. True Label for Classifying High and Low notes

From our confusion matrix, we can compute our performance metrics

Performance Metrics:

- **Accuracy**: 0.57
- **True Negative Rate**: 0.99
- **Sensitivity Recall**: 0.16
- **Precision**: 0.92
- **F1 Score**: 0.28

These results show that while the model almost never mislabels a low-pitch note as high (very high specificity), it fails to detect the majority of true high-pitch notes (low sensitivity). The high precision indicates that when it does predict "high," it is usually correct, but its recall of only 16 suggests that many high-pitch gestures remain undetected.

Across our five singers, the plot of RF LOOV AUC versus the Cosine Distance (Fig. 6) between the training set and hold-out feature importance vectors also show a clear trend:

- Amita has the lowest cosine distance (0.02) and lowest LOOV AUCs (0.525). This indicates that the model was able to measure the gestures almost perfectly but they aren't predictive of pitch for that performance.
- Prashant has the best balance of Feature Importance and AUC with a cosine distance of 0.12 and AUC of 0.667, where his gesture cues are fairly consistent between training and test (moderate stability) and delivers solid pitch-prediction accuracy.

- Anjana has the highest cosine distance of 0.15 and the second lowest AUC of 0.625 which indicates that her gesture–pitch mappings are the most idiosyncratic of the group: the features she relies on shift substantially when moving from the pooled model to her individual data, leading to less stable importance rankings and poorer singer-independent generalization.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2018.

[2] Martin Clayton et al. Hindustani raga and singer classification via 2d/3d pose estimation. *Transactions of the International Society for Music Information Retrieval*, 2024.

[3] Jane W. Davidson. Visual perception of performance manner in conductors' gestures. *Music Perception*, 24(5):433–454, 2007.

[4] Rolf Inge Godøy and Marc Leman. *Musical Gestures: Sound, Movement and Meaning*. Routledge, 2009.

[5] Anthony Gritten and Elaine King. *Music and Gesture*. Ashgate, 2011.

[6] Kelly Jakubowski et al. Conducting gesture recognition with skeleton tracking. In *Proceedings of ISMIR*, 2017.

[7] Laura Leante. Embodied knowledge and musical gesture: Perspectives from carnatic music. In Rolf Inge Godøy and Mark Leman, editors, *Music, Gesture, and Embodiment*, pages 45–60. Routledge, Abingdon, UK, 2009.

[8] Cristiano Lugaresi, Katherine McClanahan, Yanzhao Yang, Patrick Chan, Michael Hesse, Shang-Wen Ni, Tom Masters, David Portillo, Claire Crystal, Harold Steck, Chris DeChant, Jesse Egbert, Neeraj Khurana, Tyler Aming, Olga Koss, Ryan Wang, Veronica Zanella, and Praneet Mai. Mediapipe: A framework for building perception pipelines. Google Research Blog, 2020.

[9] Brian McFee et al. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, 2015.

[10] Lara Pearson. Embodied gesture and meaning in carnatic vocal performance: An ethnographic study. *Ethnomusicology Forum*, 22(2):175–192, 2013.

[11] Lara Pearson and Wim Pouw. Gesture–vocal coupling in karnatak music performance: A neuro–bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences*, 1515(1):219–236, 2022.

[12] Cynthia J. Tsay. Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, 110(36):14580–14585, 2013.

[13] V. Vaidyanathapuram Krishnan, N. Alben, A. Nair, and N. Condit-Schultz. Sanidha: A studio quality multi-modal dataset for carnatic music. arXiv preprint arXiv:2501.06959, 2025.