Supplementary information

(Exploring the Correspondence of Melodic Contour with Gesture in Raga Alap Singing)

1. Data acquisition details

Apart from the available 3-singer dataset [1], we recorded alap data across 8 new singers and used the newly augmented dataset in this work. The data is recorded with 3 cameras however this study uses only analysis of front camera. The recording setup used for the new data is given below

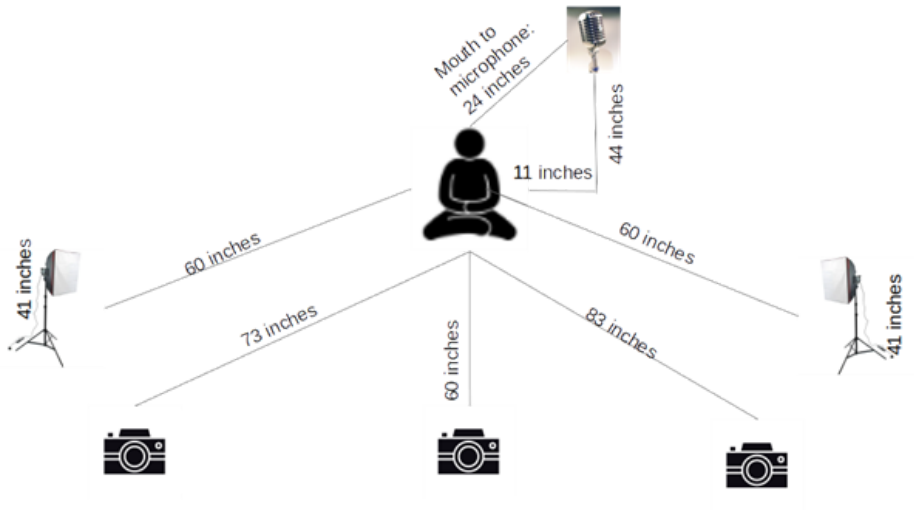

Fig 1: A representation of the setup (not drawn to scale). The singer is in a sitting position on the ground in a cross-legged fashion. There are 3 cameras, two lights and one microphone.

Audio Metadata

1. File Type : Wave (.wav)
2. Sample Rate : 48.000 kHz
3. Bit Depth : 24 bit
4. Audio Codec : AAC

Microphone Model - NEUMANN KMS 105

Video Metadata

1. Format : MP4
2. Resolution : 1920 x 1080
3. Frame Rate : 24 fps
4. Video Codec : H.264 High L4.0

Camera Model : SONY NXCAM HXR-NX5R

Singer and Raga Information

The following are the details of the singers (added in this work) using their initials.

Male Singers: AK, MG, MP, NM
Female Singers: AP, RV, SM, SS

All of the singers are **right handed** (self declared). The singers in the previous dataset [1] were AG, SCh (female) and CC (male).

Each singer sang 2 Alaps and 1 Pakad for each of the following ragas. Some of the ragas are abbreviated as in parenthesis. The ragas are the same as in [1].

Bageshree (Bag)
Bahar
Bilaskhani Todi (Bilas)
Jaunpuri (Jaun)
Kedar
Marwa
Miyan ki Malhar (MM)
Nand
Shree

2. Obtaining gesture kinematic variables from video

Our processing is similar to the methodology published in [7]. We extract 11 upper body keypoints via Openpose [5] from each frame of the video. We thus convert keypoint position coordinates for a joint into a time series. We exclude keypoint position samples extracted with a confidence of less than 0.3 and fill in the gaps by linearly interpolating using neighbouring frames. We use a Savitzky-Golay filter [6] of 4th order with 13 point smoothing window to remove jitter. We normalize the position time series for every joint based on the mean and standard deviation for that joint [z-normalization].

Additionally, in this work, we compute the velocities and acceleration profiles of the joints of interest. A smoothened derivative is computed on the 2d wrist position time series using convolution with a biphasic filter impulse response as below.

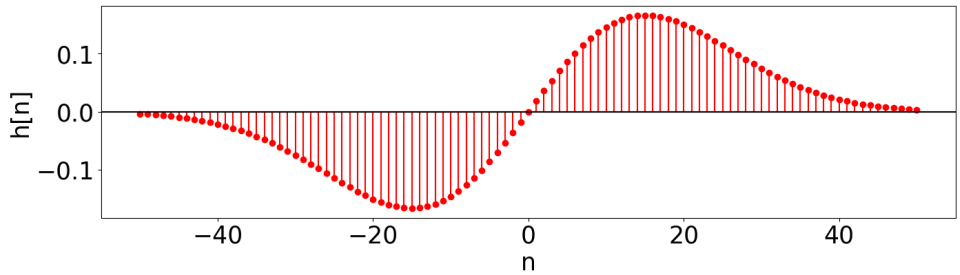

Fig 2 : Plot of impulse response a 101-point bi-phasic filter with parameters $\tau_1 = \tau_2$=15 and $d_1 = d_2 = 2$. The sampling rate here is 100 Hz.

The biphasic filter [2,3] is defined by

$$h[n] = \frac{1}{\tau_1\sqrt{2\pi}} e^{-\frac{(n-d_1)^2}{\tau_1^2}} - \frac{1}{\tau_2\sqrt{2\pi}} e^{-\frac{(n-d_2)^2}{2\tau_2^2}}$$

We convolve the biphasic filter for differentiation of the position coordinate along each axis (x and y) for each joint to obtain the corresponding velocity. We take the Euclidean norm of the individual components of velocity and refer to $v[n]$ as velocity for the joint in the rest of the work.

$$v_x[n] = h[n] * p_x[n]$$

$$v_y[n] = h[n] * p_y[n]$$

$$v[n] = \sqrt{\left(v_x^2[n] + v_y^2[n]\right)}$$

where $p_x[n]$ and $p_y[n]$ are the co-ordinates of a joint along x and y axes respectively and "*" indicates the convolution operation.

We use the same biphasic filter for differentiation of the velocity along each axis to obtain the acceleration along that axis. We take the Euclidean norm of the individual components of acceleration and refer to $a[n]$ as acceleration for the joint in the rest of the work.

$$a_x[n] = h[n] * v_x[n]$$

$$a_y[n] = h[n] * v_y[n]$$

$$a[n] = \sqrt{\left(a_x^2[n] + a_y^2[n]\right)}$$

Fig 3: Kinematic features for left wrist for a particular AV segment. X axis is in time (s) of the original alap of which this is a part. **First plot:** X-components of position, velocity, acceleration for the left wrist. **Second plot:** Y-components of position, velocity, acceleration for the left wrist. **Third plot**: Magnitude of velocity (speed) of the left wrist computed as the Euclidean norm of the vector velocity having both x and y components. **Fourth plot:** Magnitude of acceleration of the left wrist computed as the Euclidean norm of the vector acceleration having both x and y components. An identical processing is done for the right wrist.

3. Data characteristics

3.1. Stable note and non-steady region duration distributions (overall)

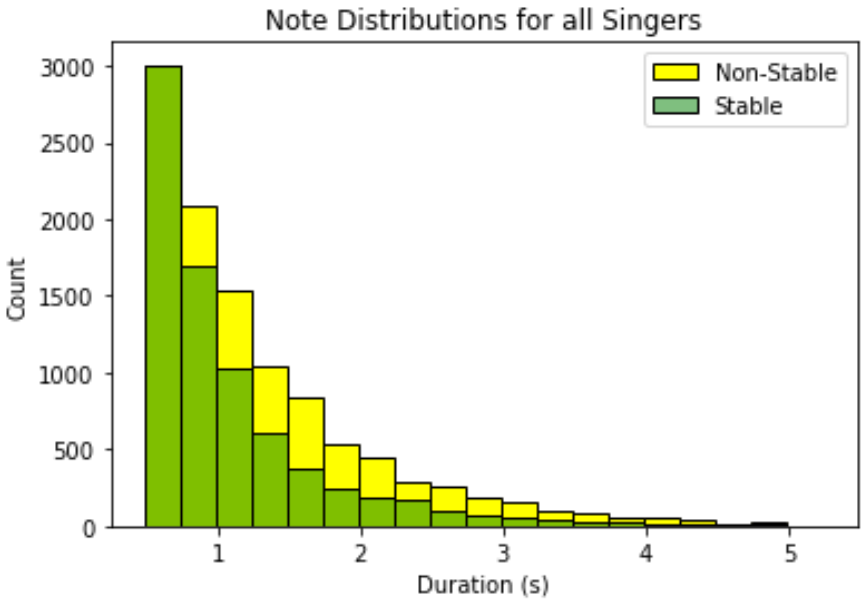

Fig 4a: Distribution of Stable and Non-stable notes with time across all singers and all ragas.

3.2. Singerwise note duration distribution

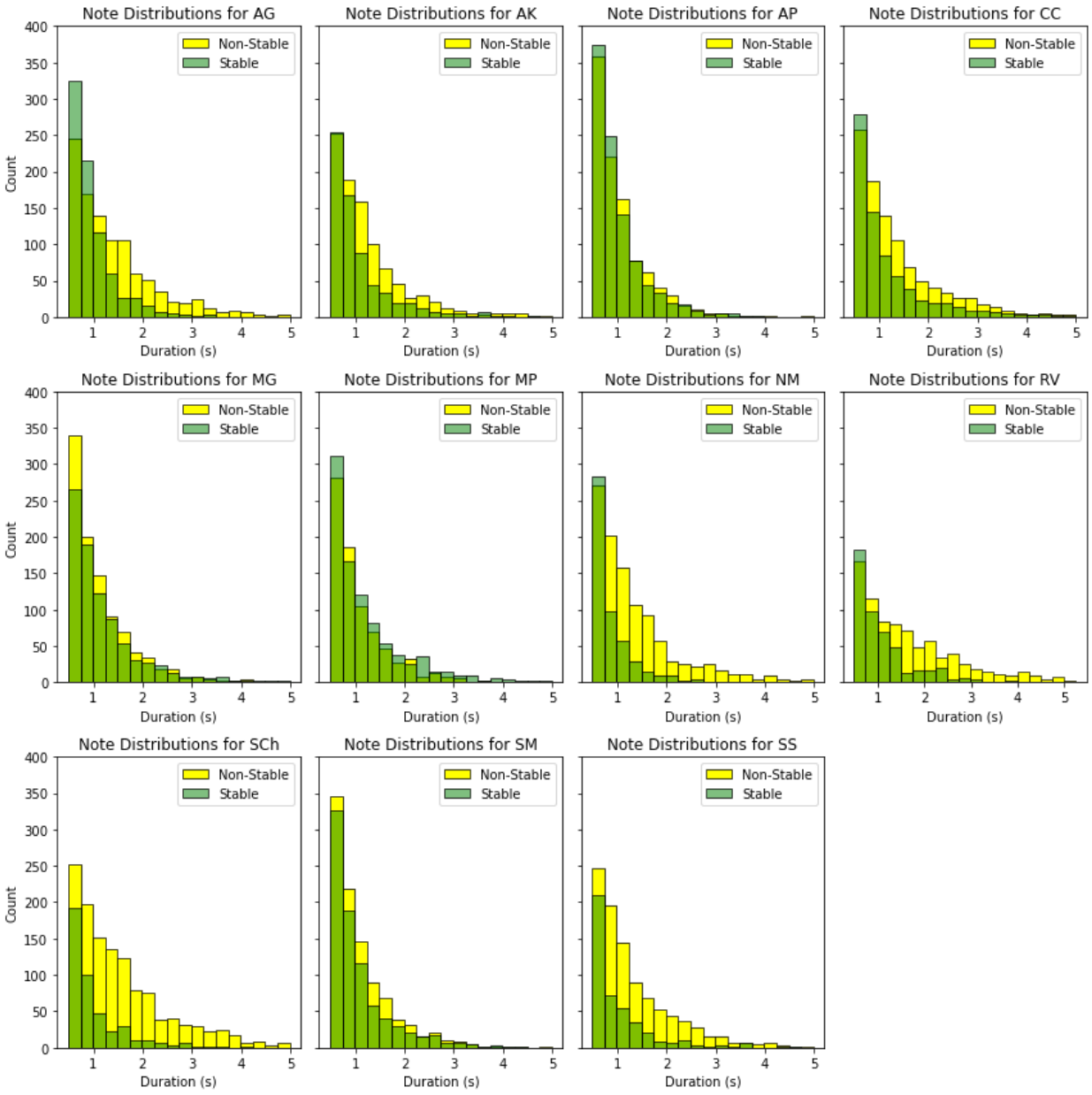

Fig 4b: Distribution of Stable and Non-stable notes with time per singer across all ragas.

3.3. Ragawise note duration distribution

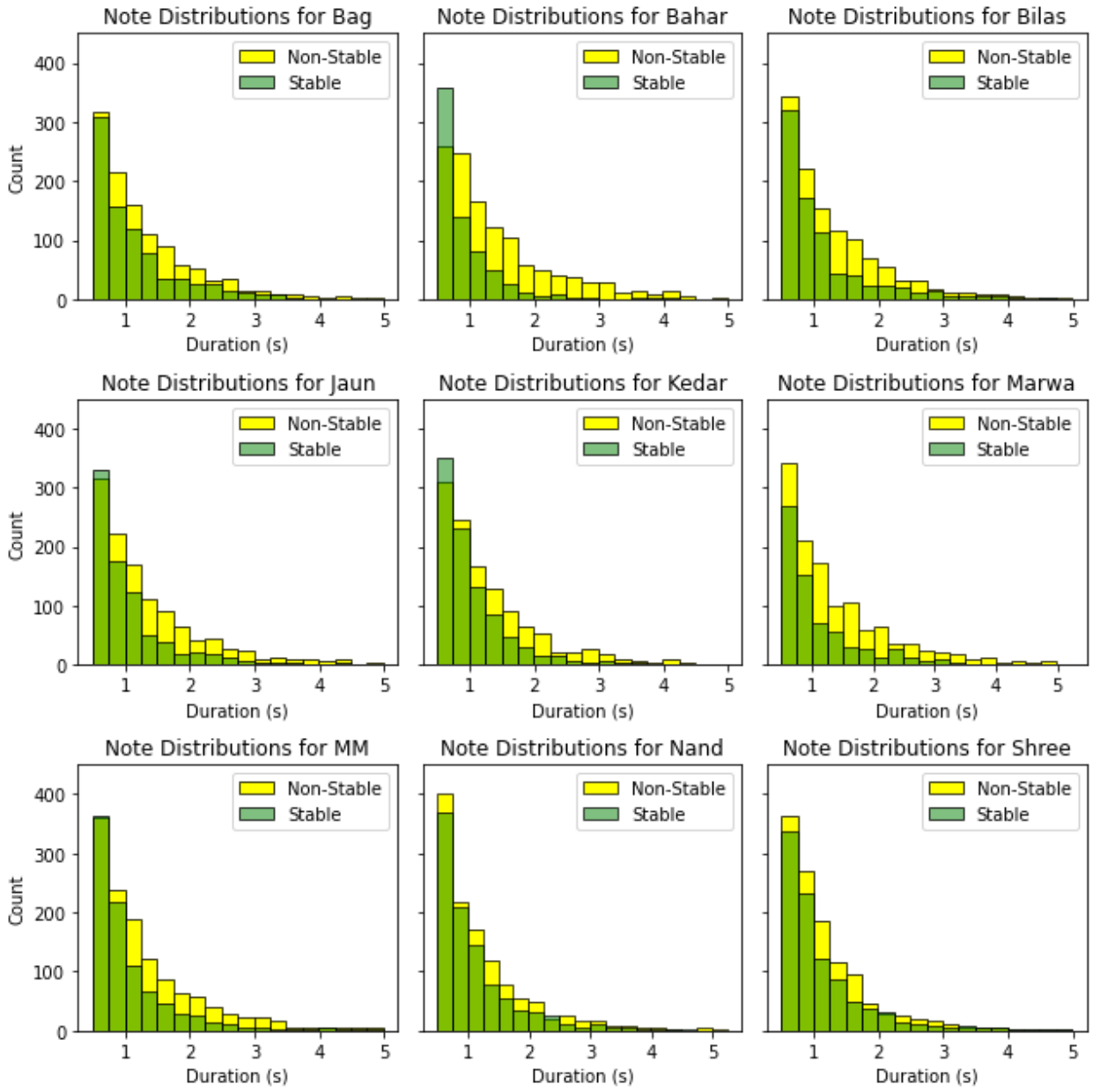

Fig 4c: Distribution of Stable and Non-stable notes with time per raga across all singers.

4. Stable note classification: Feature and classifier details

4.1. Boxplots showing feature distributions:

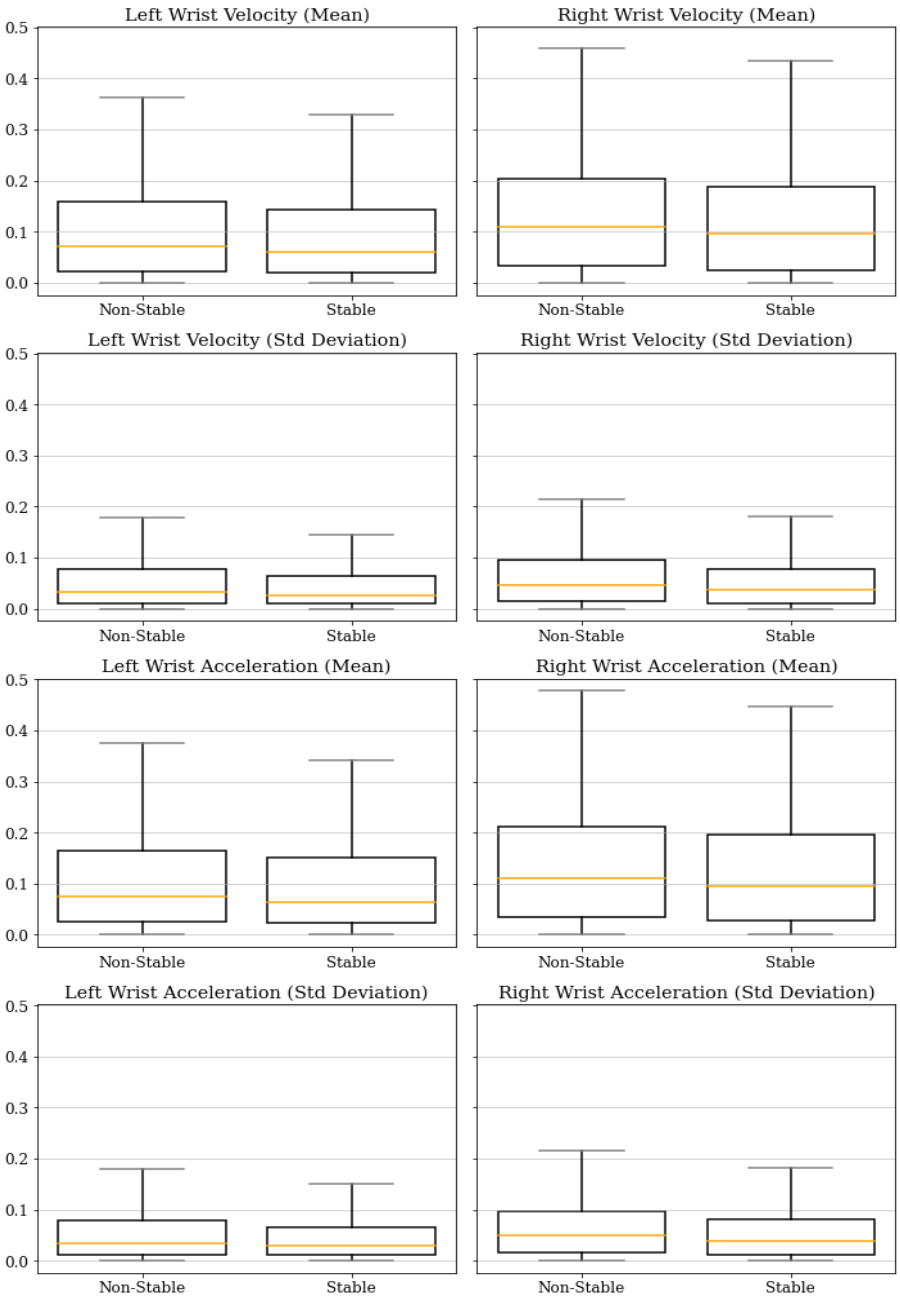

Fig 5: Boxplot of mean and std. deviation of velocity and acceleration for both wrists. These features were used in training the SVMs for overall dataset based stable-note detection.

4.2. Ragawise Classifier Performance:

Just like the singer wise SVM models, we trained models on raga wise datasets of stable and non-stable segments. As we see, the results do not show variability with raga. Each raga shows a performance similar to that of the all-singer experiment (65.7%). This indicates singer dependence is greater than raga dependence for the gestures.

Raga	**Bag**	**Bahar**	**Bilas**	**Jaun**	**Kedar**	**Marwa**	**MM**	**Nand**	**Shree**
Count	2348	2024	2331	2174	2487	2012	2581	2406	2534
%Stable	38.6	38.5	38.2	38.9	40.4	36.4	38.5	41.9	38.2
F1 score	63.4	69.8	64.5	66.0	69.6	62.5	67.3	65.7	63.4

5. Results for Raga Phrase Classification

	gmD - DTW_LR (2).			r/P - DTW_IND (8)			P\R - DTW_IND (8)		
Singer	**Like**	**Total**	**Acc**	**Like**	**Total**	**Acc**	**Like**	**Total**	**Acc**
AG	112	165	55.8	91	185	50.8	124	217	40.1
AP	93	171	50.3	79	196	59.7	69	201	65.2
CC	58	147	56.5	94	228	58.8	59	163	66.3
MG	71	140	45.7	74	160	53.8	36	180	73.9
MP	75	145	54.5	81	195	58.5	50	188	73.9
NM	67	148	39.9	85	220	61.4	47	221	76.9
RV	88	179	58.1	130	271	52	96	205	55.6
SCh	87	144	52.1	90	211	57.3	100	271	64.6
SM	113	179	50.3	128	184	30.4	82	181	58.6
SS	89	192	55.2	88	203	56.7	82	201	57.2
Overall	**944**	**1771**	**52.4**	**1035**	**2303**	**56.1**	**817**	**2157**	**65.2**

Table 4S.1 : Count of “Like” phrases, total count and accuracy (%) per singer for gesture based phrase classification. Details for only the best model as identified in Table 4 in the main paper are shown here.

Comments:- The distribution is relatively uniform across singers except the case of SM for r/P, and AG in P\R. More discussion and examples of the singer AG for P\R are presented with the accompanying videos

References:-

[1] M. Clayton, J. Li, A. R. Clarke, M. Weinzierl,L. Leante, and S. Tarsitani, “Hindustani raga and singer classification using pose estimation,” 2021. [Online]. Available: https://doi.org/10.17605/OSF.IO/T5BWA

[2] D. J. Hermes, “Vowel-onset detection,” The Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 866–873, 1990.

[3] P. Rao, T. P. Vinutha, and M. A. Rohit, “Structural segmentation of alap in dhrupad vocal concerts,” Transactions of the International Society for Music Information Retrieval, vol. 3, no. 1, 2020.

[4] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi et al., “Kerastuner,” https://github.com/keras-team/keras-tuner, 2019

[5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 43, no. 1, pp. 172–186, 2021.

[6] Press, William H., and Saul A. Teukolsky. "Savitzky-Golay smoothing filters." *Computers in Physics* 4.6 (1990): 669-672.

[7] Clarke, A., Weinzierl, M., & Li, J. (2021). Pose estimation for Raga (Version v1.0.1) [Computer software]. https://github.com/DurhamARC/raga-pose-estimation