Multimodal Raga Classification from Vocal Performances with Disentanglement and Contrastive Loss

Supplementary Material

1. Hyperparameters

We present the hyperparameters for the best models and each split in this section.

a) **Audio Models.** The hyperparameters are for F0+VM+GR and correspond to the Table 7 of the manuscript.

Parameter	Values	Split 1	Split 2	Split 3
Temporal resolution	10ms, 20ms	20 ms	10 ms	10ms
No. Conv layers	1,2	2	1	2
No. Conv filters	4,8,16,32,64,128	4,8	8	4,4
Kernel Size	3,5,7	5,7	7	3,3
Inception filters (n11)	4,8,16,32,64,128	4	8	4
Inception filters (n21)	4,8,16,32,64,128	8	8	8
Inception filters (n31)	4,8,16,32,64,128	8	32	8
Inception filters (n32)	4,8,16,32,64,128	8	32	32
Inception filters (n41)	4,8,16,32,64,128	32	64	128
Inception filters (n42)	4,8,16,32,64,128	128	64	128
Inception filters (n43)	4,8,16,32,64,128	64	128	64
Regularization (L2) weight	0-1e-4	0.002	2*1e-4	0.004
Dropout Rate	0-0.5	0.26	0.12	0.45
Learning Rate	0.01,0.001,1e-4	0.01	0.01	0.001

Table 7a: Best hyperparameters per split for audio F0+VM+GR corresponding to Table 7 in manuscript. Hyperparameters for split 3 gives best average performance and is used for experiments whose results are reported in Table 9. The inception filter ids n11-n43 correspond to Figure 4.

b) **Video Models.** The hyperparameters are for PVA-WE+GR and correspond to the Table 7 of the manuscript.

Parameter	Values	Split 1	Split 2	Split 3
Temporal resolution	10ms, 20ms	10 ms	10 ms	10ms
No. Conv layers	1,2	2	2	2
No. Conv filters	4,8,16,32,64,128	4,4	4,4	4,4
Kernel Size	3,5,7	3,5	3,3	3,3
Inception filters (n11)	4,8,16,32,64,128	4	4	4
Inception filters (n21)	4,8,16,32,64,128	8	4	8
Inception filters (n31)	4,8,16,32,64,128	8	8	8
Inception filters (n32)	4,8,16,32,64,128	16	8	8
Inception filters (n41)	4,8,16,32,64,128	32	32	32
Inception filters (n42)	4,8,16,32,64,128	32	32	64
Inception filters (n43)	4,8,16,32,64,128	64	32	32
Regularization (L2) weight	0-1e-4	0.001	0.003	0.003
Dropout Rate	0-0.5	0.17	0.32	0.30
Learning Rate	0.01,0.001,1e-4	0.01	0.1	0.01

Table 7b: Best hyperparameters per split for audio PVA-WE+GR corresponding to Table 7 in manuscript. Hyperparameters for split 1 gives best average performance and is used for experiments whose results are reported in Table 9. The inception filter ids n11-n43 correspond to Figure 4.

c) Multimodal Fusion Model (BPCL)

Parameter	Values	Split 1	Split 2	Split 3
Common Embed Dimension	2-128	128	118	115
Temperature	0-1	0.48	0.16	0.71
Regularization L2 weight	0-1e-4	0.001	0.001	0.002
Dropout Rate	0-0.5	0.27	0.13	0.12
Learning Rate	0.01,0.001,1e-4	1e-4	0.001	0.001

Table 8a: Best hyperparameters per split for multimodal model BPCL corresponding to Table 8 in manuscript. Hyperparameters for split 1 gives best average performance and is used for experiments whose results are reported in Table 9.

2. Effect of silence on classification accuracies

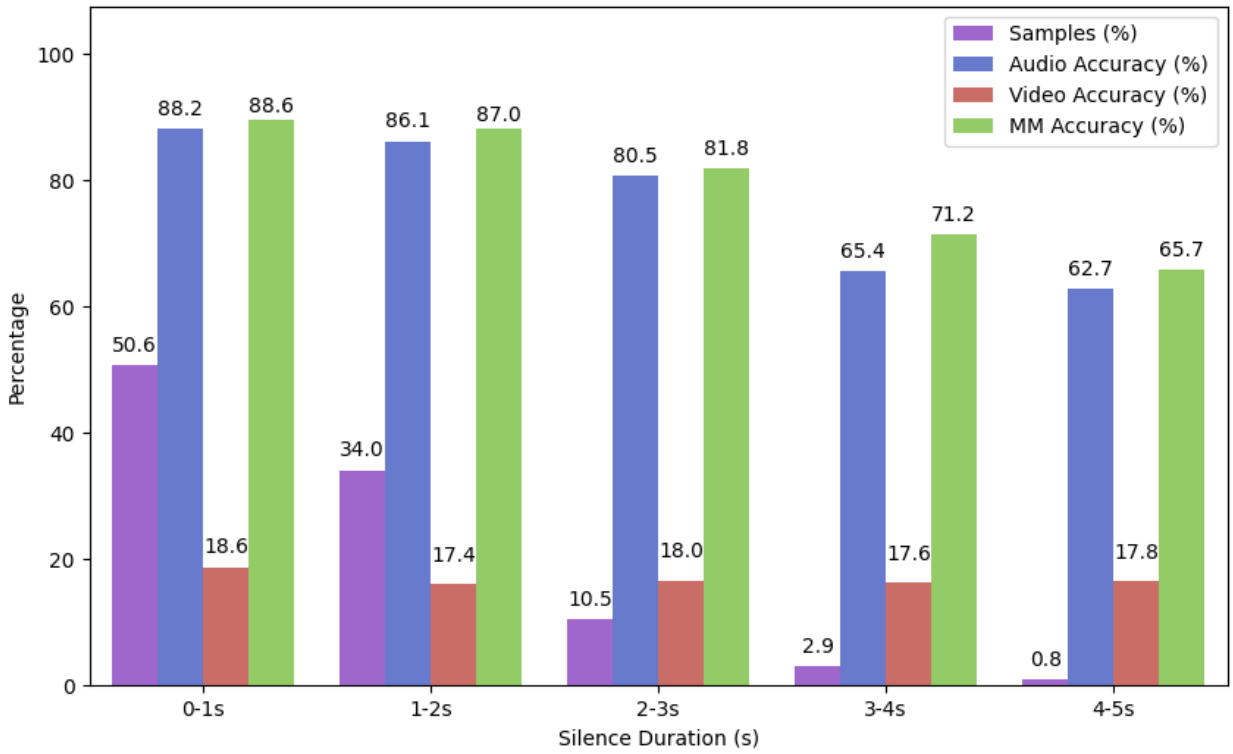

Figure 1 : Percentage of samples and accuracy (%) across multiple modalities for validation data on split 1-3 shown by the total duration of silence in a 12s snip. 98.8% of validation data (12055 samples) have total silence less than 5s

Table 10 of the manuscript shows the classification accuracies split by less than 2s and more than 2s. Figure 1 above show a more fine-grained split by the total duration of silence. The number of samples with higher duration of silence reduces. We observe that the audio accuracy falls consistently as the silence duration increases. We can see that the video accuracy though poor remains roughly constant for different durations of silence, and this leads to better improvement in multimodal accuracy (over audio accuracy) when the silence is higher. The relatively constant video accuracy is because singers continue their gesture even when not vocalizing. This shows the importance of the video modality despite the relatively poor accuracies.

3. TSNE Plots

We consider the features at the output of inception layers for each of the classification tasks of singer and raga identification and inspect the t-distributed Stochastic Neighbour Embedding (TSNE) [1]..

Figures 2a and 2b shows the TSNE features for singer classification for Split 1 using audio and video features respectively. The instances are colour coded by ground-truth singer labels. While the audio features look diffused, the video features are well clustered consistent with high classification accuracy in Table 6 of the manuscript.

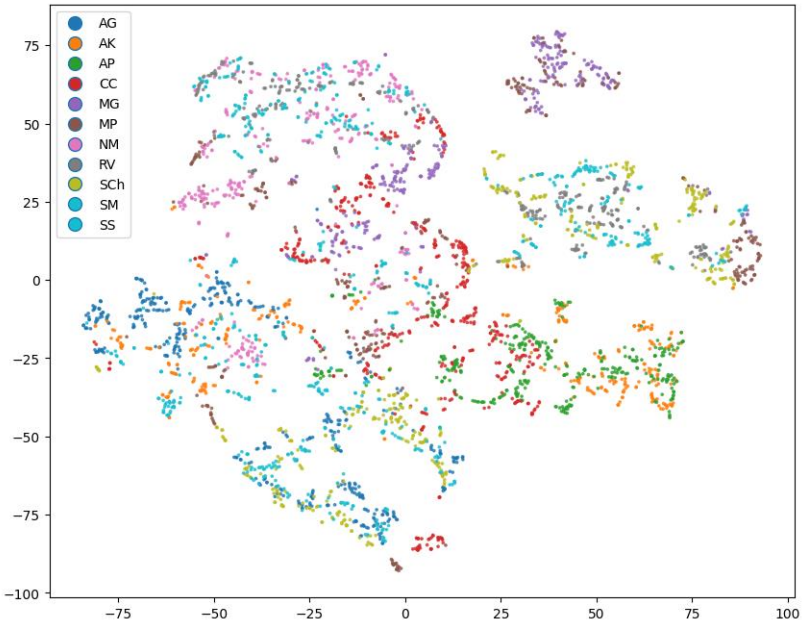

Fig 2a: TSNE plot on extracted features from singer classification models based on audio for Split 1 validation data. We use the "F0+VM" model for audio and "PVA-WE" for video. The accuracies are reported in Table 6 of manuscript.

[1] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.11 (2008)

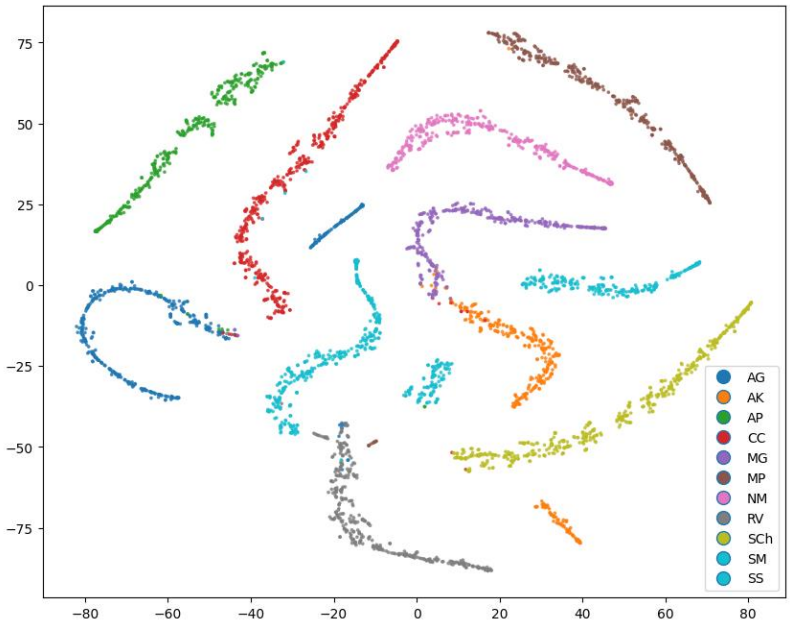

Fig 2b: TSNE plot on extracted features from singer classification models based on video for Split 1 validation data. We use the "PVA-WE" for video. The accuracies are reported in Table 6 of manuscript.

Figures 3a and 3b shows the TSNE plots for raga classification of Split 1 using audio and video features respectively. The points are colour coded by the ground-truth raga labels.

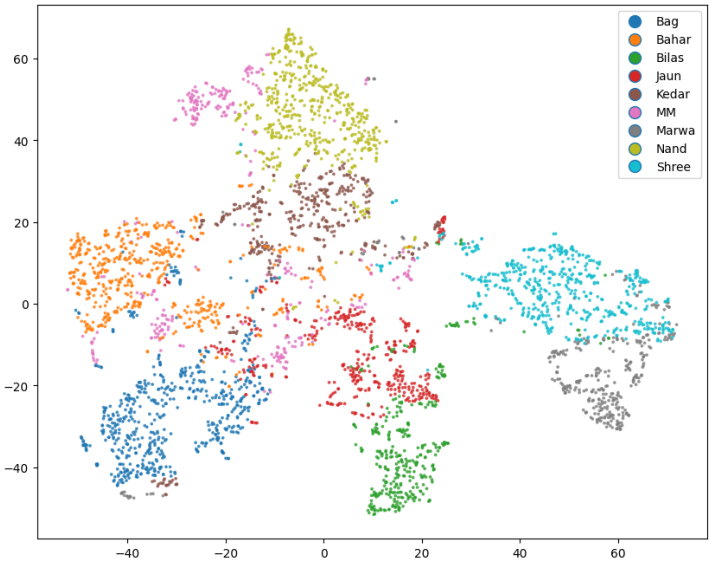

Fig 3a TSNE plots on extracted features from raga classification models based on audio for Split 1 validation data. We use the "F0+VM+GR" model for audio. The accuracies are reported in Table 7 of manuscript.

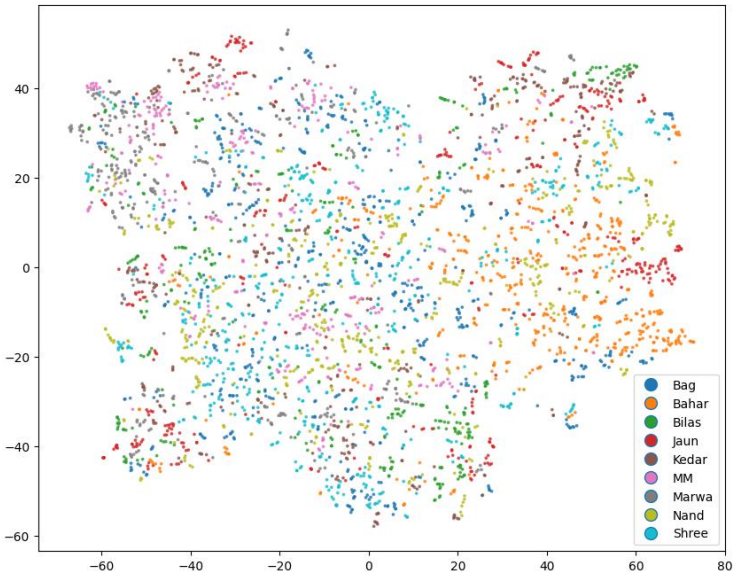

Fig 3a TSNE plots on extracted features from raga classification models based on video for Split 1 validation data. We use the "PVA-WE+GR" model for video. The accuracies are reported in Table 7 of manuscript.

We note the clustering of raga labels with audio but not so with video features - which is reflected in the raga classification accuracy metrics. Also, we observe that some of the observed misclassifications of the audio e.g. MM and Bahar, Kedar and MM etc. are indeed closer together in the TSNE projected space whereas Bageshree and Shree which show the least errors in the audio confusion matrix in Figure 9 of the manuscript are well separated in the TSNE plot.