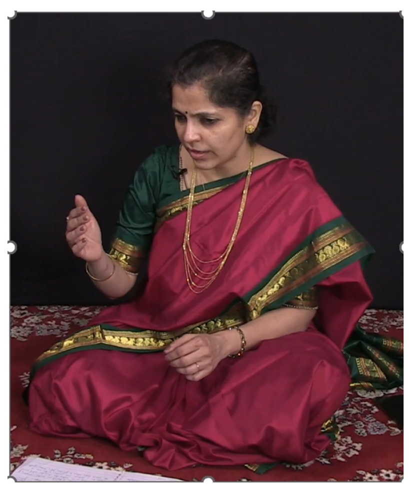
Raga Classification from Vocal Performances Using Multimodal Analysis

Martin Clayton[1], Preeti Rao[2], Nithya Sikharpur[2], Sujoy Roychowdhury[2] , Jin Li[1]
[1]Department of Music, Durham University, United Kingdom
[2]Department of Electrical Engineering, Indian Institute of Technology Bombay, India

Motivation



A singer's gesture while performing a raga

- Musical gesture studies [1] build on those of speech-accompanying gesture
- Indian vocal music has featured prominently in this field due to its rich use of manual gestures
- Gesture styles are idiosyncratic (varies between teacher-student or siblings)
- Nonetheless, gestures often seem to accompany (and illustrate) aspects of melody
- *"Can ragas be classified using movement information alone, or can movement information help to disambiguate raga identity?"*

Multimodal Classification Objective

Train a deep learning classifier using

- Unimodal features from audio and video
- Try different multimodal classification methods and compare performance

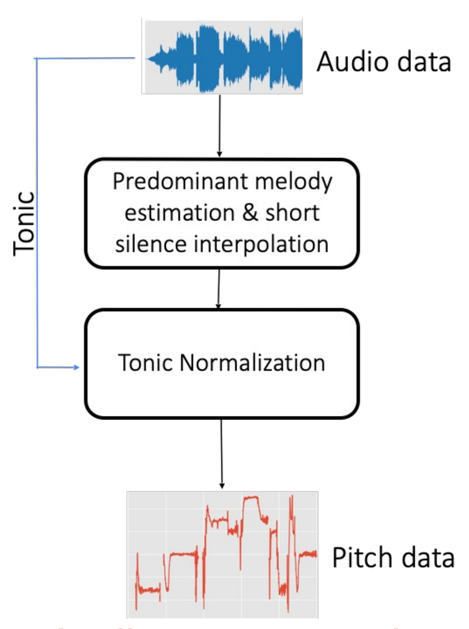
Dataset Description and Raga Information

We use an OSF dataset comprising alap (2 takes/(raga, singer) x 3 mins/take) and characteristic phrases (pakad, 4 takes/raga x 20s/take) in 9 ragas, sung by 3 professional artists (~3.5 hours total)

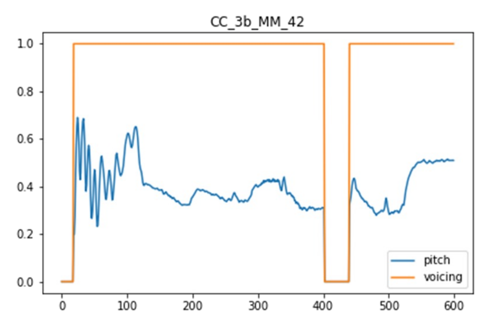
Raga	Scale
Bageshree (Bag)	S R g m P D n
Bahar	S R g m P D n N
Bilaskhani Todi (Bilas)	S r g m P d n
Jaunpuri (Jaun)	S R g m P d n
Kedar	S R G m M P D N
Marwa	S r G M D N
Miyan ki Malhar (MM)	S R g m P D n N
Nand	S R G m M P D N
Shree	S r G M P d N

The notes of each raga; S is the tonic note

Extracting Audio Features



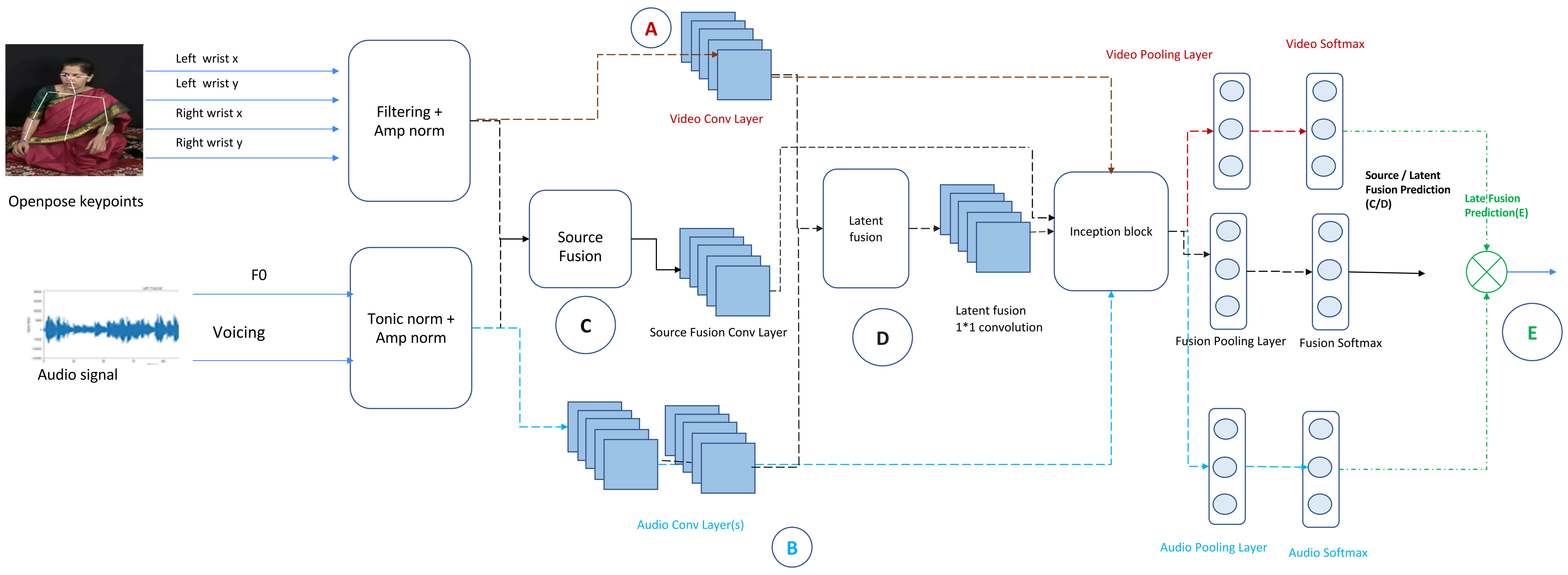
Audio preprocessing



Sample audio pitch and voicing mask

- We use source separation of vocal from background tanpura.
- We extract vocal pitch and binary voicing at 10 ms intervals. Interpolate across short silence segments (<250 ms)
- We apply tonic normalization to obtain pitch in cents with reference to the singer's tonic.

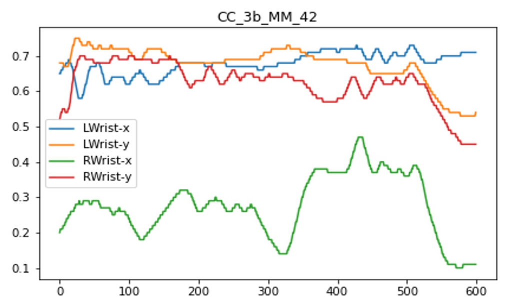
Multimodal Classification Architecture



Multimodal raga classification architecture

Extracting Video Features

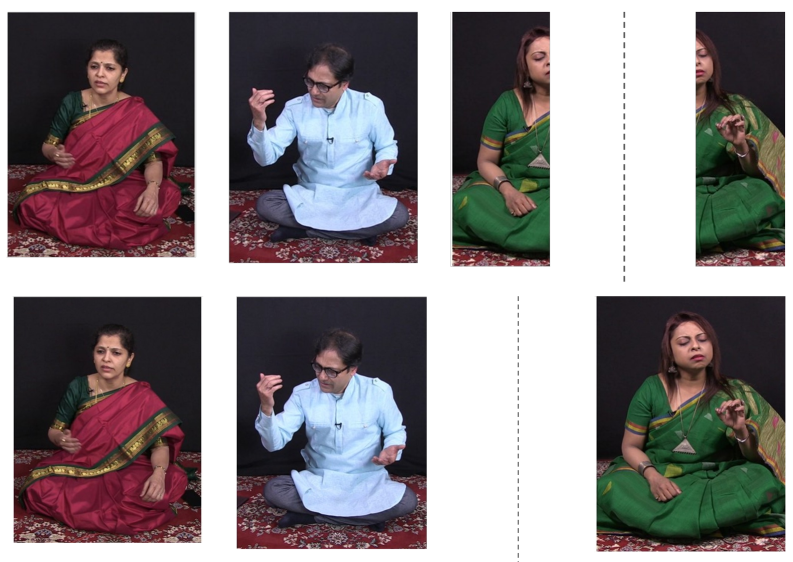
Sample OpenPose output & normalization box



Sample video features for one 12s clip

- We use OpenPose 2D pose estimation [2] to track a set of upper-body keypoints
- Normalize keypoints based on square bounding box around singer to a range of [0,1]
- Use only the positions of right and left wrists
- Any missing data is interpolated and lowpass filtering applied for smoothing

Train/Test Splits



Data split	Seen split			Unseen split		
	AG	CC	SCh	AG	CC	SCh
Train	5590	5487	5588	4715	4105	4304
Validation	972	1075	974	1847	2457	2258

Seen and unseen split sizes by 12s segments

- Recordings are split into overlapping 12s segments (order of phrase duration)
- Seen singer split - 1 alap take of one singer in validation, rest in train
- Unseen singer split - All recordings of one singer in the validation, rest in train

Architecture Components

- Convolutional layers for feature extraction
- Inception layer to process multiscale information
- Layers separately hyperparameter tuned for individual experiments

Unimodal Experiments

- Unimodal Video- Use video features (model A)
- Unimodal Audio- Use audio features (model B).

Multimodal Experiments

- Source Fusion – Combine audio and video features (model C)
- Latent Fusion – Use frozen weights of the best model from (A) and (B) and train inception and final layers (model D)
- Late fusion – Train a classifier (RF etc.) on top of predicted softmax of unimodal models A and B (model E)

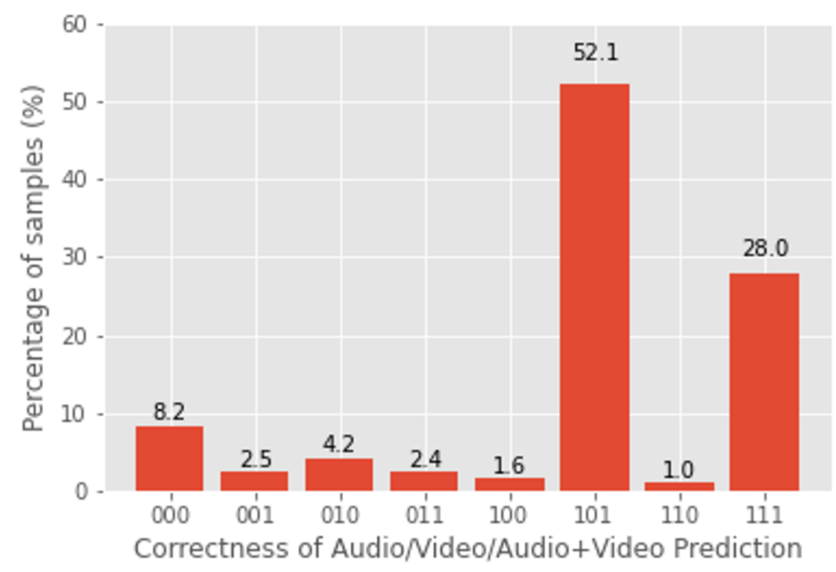
Model Results

Data Split	Seen Singer		Unseen Singer	
	Audio	Video	Audio	Video
AG	92.1	36.3	76.9	14.3
CC	79.4	31.8	60.4	13.8
SCh	77.0	39.2	67.2	10.0

Unimodal validation accuracy

Model Type	Model Name	AG	CC	SCh	Mean
A	Video	36.3	31.8	39.2	35.8
B	Audio	92.1	79.4	77.0	82.8
C	Source fusion	30.1	42.4	35.8	36.1
D	Latent fusion	**93.3**	**82.7**	**79.2**	**85.1**
E1	Equal voting	85.9	73.7	67.9	75.8
E2	Stacking classifier – RF	81.9	74.2	76.3	77.5

Multimodal validation accuracy



Comparison of unimodal and latent fusion models. 3-bit code indicating if audio, video and latent fusion models are correct (1) or wrong (0) respectively

Conclusions

- Unimodal audio results much better than video
- There is complementary information in video to improve multimodal performance over audio with latent fusion

References

[1] Godøy, R. Inge, M. Leman, eds. Musical gestures: Sound, movement, and meaning. Routledge, 2010.

[2] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Scan accompanying QR code for supplementary material.