EXPLORING THE CORRESPONDENCE OF MELODIC CONTOUR WITH GESTURE IN RAGA ALAP SINGING

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in
submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

Musicology research suggests a correspondence between manual gesture and melodic contour in raga performance. Computational tools such as pose estimation from video and time series pattern matching potentially facilitate larger-scale studies of gesture and audio correspondence. We present a dataset of audiovisual recordings of Hindustani vocal music comprising 9 ragas sung by 11 expert performers. With the automatic segmentation of the audiovisual time series based on analyses of the extracted F0 contour, we study whether melodic similarity implies gesture similarity. Our results indicate that specific representations of gesture kinematics can predict high-level melodic features such as held notes and raga-characteristic motifs significantly better than chance.

1. INTRODUCTION

Manual gesturing by singers is an integral part of vocal music performances in the Indian classical traditions. Previous work has demonstrated that singers' gestures have several different referents and functions: for example, they may relate to the rhythmic structure of the music (marking a steady beat or tala cycle) or play a role in signalling to co-performers or audience members, as well as appearing to accompany or illustrate aspects of the melody being sung. In the latter case, hand movements sometimes appear to correspond to pitch height (i.e. ascending pitch co-occurs with one or both hands rising and/or moving to one side); at other times they relate to other aspects of melody, such as the tension felt while sustaining certain notes, or the image or abstract design visualised by the performer [1–5].

Little computational work has been carried out on gesture-to-audio correspondence in Hindustani vocal music. Paschalidou [6] carried out research on a motion capture dataset of solo alap recordings in the dhrupad genre, looking at a range of movement and audio features in relation to the concept of 'effort': although she found correspondences, generalising across performers proved challenging.

Clayton et al [7] explored the use of movement data to classify 12-sec excerpts drawn from a corpus comprising 3 singers performing 9 common Hindustani ragas in the khyal genre. The use of solo alap meant the gestures cannot refer to either metric structure or interaction with co-performers, and thus relate predominantly to the melody of the ragas being presented. An inception block preceded by independently trained convolution layers for each of audio and gesture time series classification provided the best performance in the context of singer-dependent raga classification, especially reducing the confusion between melodically similar ragas with respect to the otherwise high-performing audio-only classification. While the work demonstrated the complementarity of gesture and melodic profiles in relation to raga identity, we are more interested in the present work in understanding which characteristics of gesture correlate with specific melodic characteristics. Further, given that the dataset of [8] was limited to 3 singers and therefore not suited to cross-singer studies, we present here a considerably enlarged corpus with 8 additional performers, collected following a similar methodology, as summarised in Table 1.

Dataset	Singers	Ragas	Pakad	Alap	Dur(min)
Study in [7]	3 (1M,2F)	9	37	55	193
Current Work	11(5M,6F)	9	109	199	664

Table 1: A summary of the newly augmented audiovisual dataset compared with that of closest previous work [7].

In the related Karnatak tradition, Pearson's research has looked at the role of gesture in vocal teaching [9]. The relation between the acoustics and the kinematics was studied in recent work by Pearson & Pouw using the tracking of left and right wrist positions [10]. They manually segmented the gesture tracks and studied the correspondence of various kinematic extrema with the temporally aligned changes in the acoustics (fundamental frequency, or F0, and amplitude envelope). A correspondence was established between the magnitudes of local peaks in acceleration and changes in F0, in line with previous work in co-speech gesturing [11].

In this work, we study the newly expanded corpus of solo alap recordings. Since the same set of 9 ragas is performed by all the singers (11 in this case), we can explore commonalities in the gestures used by different singers for

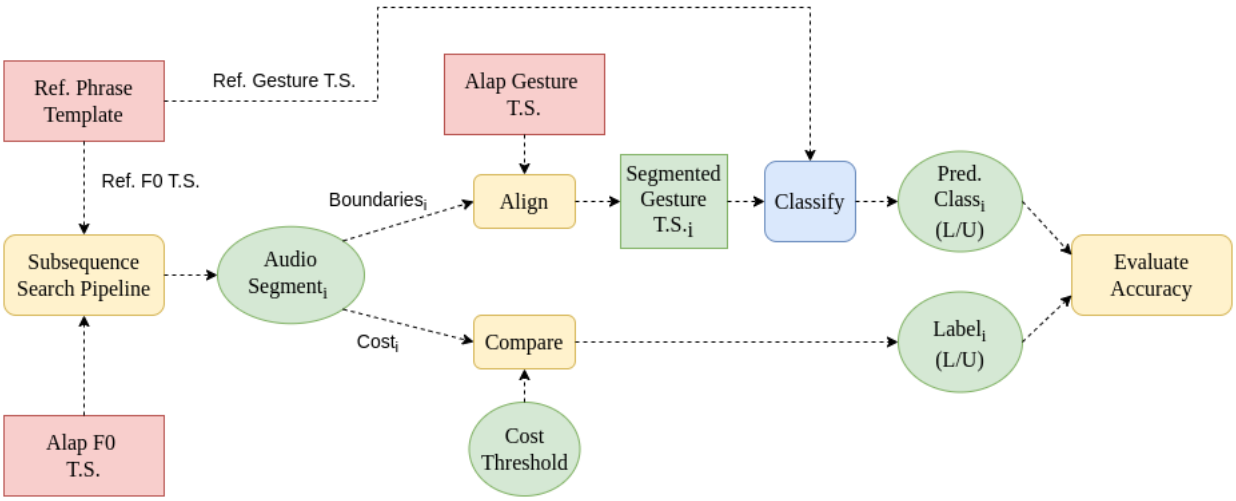

Figure 1: The overall testing and evaluation framework for the raga phrase-based segmentation. The audio and visual components of a candidate AV segment (i-th segment from an alap) are separately compared with the respective audio and visual components of a reference phrase segment (from a pakad) to see whether they are together consistent in their estimation of similarity with the reference phrase. We note that the gesture T.S. (time series) is multidimensional while the audio T.S. is a unidimensional sequence of F0 samples.

particular raga-specific melodic movements. That is, in contrast to the body of previous work, we use musically motivated units, implied by the raga melodic structure, to group the representations of melody and gesture. The aim of the study is to investigate correspondences between the singers' movements (captured in the time series for x- and y-coordinates of their wrist positions) and the melodies they sing (represented as F0 contours).

Figure 1 depicts our overall framework. The melodic phrase segments are obtained for each alap audio via a subsequence search using a reference audio template (such as a manually labeled phrase segment). The audio segment start and end times are then used to identify the corresponding time-synchronised video segment. The audio and video segments are individually processed to compute audio-based similarity and video-based similarity with respect to the corresponding components of the AV (audio-visual) reference template. We now seek to quantify the extent to which video-based similarity predicts audio similarity. We simplify the evaluation task to comparing, across the two modalities, the following binary labels: L (i.e. close to, or Like, the reference) or U (Unlike the reference).

In the next section, we provide the details of our dataset. This is followed by a discussion of the audiovisual segmentation methods. The experiments and results are presented in the final two sections of the paper.

2. DATASET AND PREPROCESSING

We consider our dataset of vocal alap performances by 11 professional musicians performing 2 alaps each of 9 ragas. Each alap is about 3 minutes long. The singers also contributed shorter 'pakad' recordings, rendering some of the key phrases of each raga in a brief format of a few seconds. The total duration of this newly expanded dataset (summarised in Table 1) is about 11 hours. Each piece was recorded using three video cameras and separate microphone; only the central camera is used in the current analyses. While each alap is labeled only by singer and raga, we carry out further manual annotation of the pakad audio files for selected raga phrases as used in this study. That is, all the pakads of a given raga across the 11 singers are searched for instances of the desired phrase (e.g. gmD in raga Bageshri). This task, carried out by a musician, is facilitated by the fact that the pakad is almost always sung with solfege (unlike the alap).

Our audio and video processing pipelines closely follow those of [7]. An initial stage of audio suppression of the background drone is obtained via source separation [12]. This is followed by pitch and voicing detection at 10 ms intervals using monophonic pitch detection based on short-time autocorrelation analysis [13]. Brief unvoiced regions (less than 400 ms) arising from short breath pauses and consonant utterances are filled in via cubic spline interpolation to obtain the continuous pitch contours associated with melodic movements that are bounded by silence (>400 ms) on both ends. These are termed 'Silence-Delimited Segments' (SDS). The pitch contour is tonic-normalised using an automatically detected (and manually verified) tonic to obtain the F0 (cents) contour sampled at 10 ms intervals [14].

In order to extract the movement data, the central video view of each piece is processed using the OpenPose pose estimation algorithm, which generates x- and y-coordinates for 11 upper body joints [15]. We select the right and left wrist coordinates. Any missing data are interpolated and each of the time series is low-pass filtered to remove jitter. The position time-series, originally sampled at 25 fps is interpolated to 100 samples/sec to synchronise

it with the sampled F0 contour. Other important low-level human motion descriptors include velocity (rate of change of the 2d position) and acceleration (rate of change of the velocity) [16]. We derive velocity and acceleration profiles from the 2d position time-series of each joint by computing derivatives. A robust estimate of the derivative is obtained via a differencing kernel such as a biphasic filter with its controllable smoothing parameters [17, 18]. We find that a 101-point filter achieves a lowpass filtering of about 2 Hz, giving a sufficiently smooth and physiologically plausible movement acceleration profile [19]. We eventually obtain the 8-dimensional gesture time series of position (x,y), velocity and acceleration for each of the left and right wrists for each of the singer-alap and pakad recordings. In this, we include the synchronized F0 contour to get the complete audiovisual time series for an alap, now in the form of a sequence of SDS. A detailed review of the data collection and processing appears in the supplementary material.

3. SEGMENTATION METHODS

The first stage of segmentation of the synchronised AV time series comprises the silence-delimited segments (SDS) obtained in the previous section. We discard segments of duration less than 500 ms as too limited for our further analyses. The retained SDS, numbering approximately 30 per alap, have a mean duration of 5.2 s with less than 1% (of the total count of 6012) exceeding 20 s. As discussed next, we apply melodic segmentation principles to each SDS to obtain stable note and raga-characteristic melodic movements or phrases that can help us explore the links between specific musical expressions and the corresponding gestures.

In a top-down approach, the alap can be segmented into its phrases. A raga phrase, although notated simply by its solfege sequence, has a melodic-rhythmic realisation comprising specific intonations and durations of its constituent svaras, together with the transitions to/from neighbouring svaras [20]. On the other hand, in a bottom-up approach, the melodic contour can be viewed as comprising the following broad categories of segments: stable notes, and the transitions between the notes which can include distinctive melodic ornaments such as glides (meend) and oscillatory movements (andolan) apart from steep changes of pitch or pauses [21]. Figure 2 presents an example of an SDS that comprises a variety of stable and transitional sub-segments. It is therefore of interest to examine audiovisual correspondences in the context of the distinct types of melodic movements. The two different audio-based segmentation procedures are detailed next.

3.1 Stable note segmentation

To identify occurrences of stable or sustained tones, the continuous F0 contour corresponding to an SDS is searched for instances in which the same raga note (svara) is sustained for > 250 ms. That is, a stable note is defined as a region where the F0 lies within a 25 cent interval of the mean intonation of the raga note. This is based on

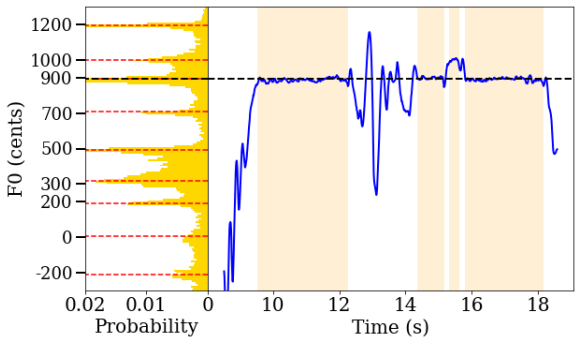

Figure 2: A sample SDS with identified steady notes (shaded regions of blue F0 contour) and pitch salience distribution (on the left) computed from the entire alap audio with detected svara locations highlighted.

past work that associated the similar duration and intonation parameters with a listener's percept of a held note [22]. Further, given that a svara may not be realised on the equitempered grid but rather with a raga-specific intonation, we use a finely binned pitch salience distribution computed across the alap to establish the svara locations [22]. Stable note regions corresponding to the same svara that are separated by less than 100 ms are next merged. The boundaries of the so detected stable notes are shown in the example of Figure 2. Across our alap dataset, stable notes were found to range from 0.25 s to 9.9 s with a mean of 0.73 s.

In a similar vein, we considered the segmentation of another characteristic melodic movement, the glide (or slide). This has been attempted previously via the quality of a linear fit to the F0 contour for Indian popular vocal music [23]. However we found that the variety and complexity of glide movements in raga music make it challenging to develop a universal glide detection algorithm. We therefore resort to template-based phrase detection for the purpose, as explained next.

3.2 Phrase-based segmentation

As depicted in Table 2, the raga motifs selected for our exploration include a distinctive upward slide of an augmented fourth in Shri, a falling slide of a fourth in Nand, and a three-note ascending phrase in Bageshri. The chosen phrases are highly characteristic of the corresponding raga and occur in the raga alap with relatively unchanged melodic shape, prompting the question about whether their gesture executions also bear some measurable similarity. The corresponding pakad phrases serve as templates for the segmentation of the alaps for the chosen raga across the 11 singers. We obtain a number of templates of the given phrase from across the 11 singers' pakads. The set of templates represents the diversity in the realization of the phrase across and within singers. This is manually reduced to a set of 6 templates per phrase while retaining the diversity. Figure 3 shows a few examples for each of the phrases chosen for the current study. We observe that the simple notation used to represent the up or down slide (/,

Raga	Svara (Notes)	Phrase
Bageshri	S R g m P D n	gmD
Shri	S r G M P d N	r/P
Nand	S R G m M P D N	P\R

Table 2: The ragas and phrases used in this study. The svaras S r R g G m M P d D n N correspond to the 12 notes of the Western chromatic scale with S representing the tonic. The symbols / and \ denote the upward and downward slide respectively [24], [25].

\) belies the complexity of contour shapes defined by raga grammar. Also clear are the essential shape features that point to the need for dynamic time warping (DTW) based comparisons [26]. Next, the following steps (also visualised in Figure 4) lead to the desired segmentation of the alap audio files for each selected raga phrase.

1. The six phrase templates from the pakads are warped to the same target length (that of the 3rd template in increasing length order in the set) using a penalty parameter that discourages large deviations from the diagonal path. This helps to ensure that the subsequence DTW matching costs can be meaningfully compared across the templates.

2. As shown in the middle panel of Figure 4, constrained DTW based subsequence search is executed on each SDS with each of the 6 warped audio templates (WAT) to obtain for each WAT the lowest cost match that satisfies a duration criterion (> 0.5s) in order to avoid cases of pathological warping [27]. Such matches are accepted as valid and stored with the cost, temporal boundaries and WAT index. In case no valid match is returned (in the top 20 retrieved responses) for a particular template, that SDS-template is not considered further. This step leaves us with between 1 to 6 best matched segments per SDS along with the associated DTW costs.

3. Next, we pick the single lowest cost for each SDS and use this value to cluster the entire set of SDS, across 22 alaps of the raga, into 2 clusters by fitting a kernel density estimate (KDE) to the distribution of costs as shown in Figure 5 [28]. The cost value coinciding with the lowest point in the valley between the peaks is used as a cost threshold to label each SDS as one of the two classes:'Like' (i.e. similar to the raga motif) and 'Unlike' (different from the raga motif). These are the labels we would like to predict from the corresponding gesture time series segments in the context of our investigation of audiovisual correspondence.

4. In order to increase the number of examples for the gesture-based prediction task, we club all the different template matches obtained in Step 2 for the same SDS under the same label. This was justified by our observation that the SDS labeled Like (L) in Step 3 typically exhibited similar low cost matches across all templates of the same phrase. The SDS marked Unlike (U), on the other hand, exhibited a relative wide spread in cost values above the threshold, similar to that depicted in Figure 5.

Finally, with the audio segments computed in this section, we extract the corresponding temporally synchronised gesture time series for each SDS and WAT pair. In the next section, we report our experiments on testing various kinematic features for the prediction of the corresponding audio-derived labels in our two distinct tasks.

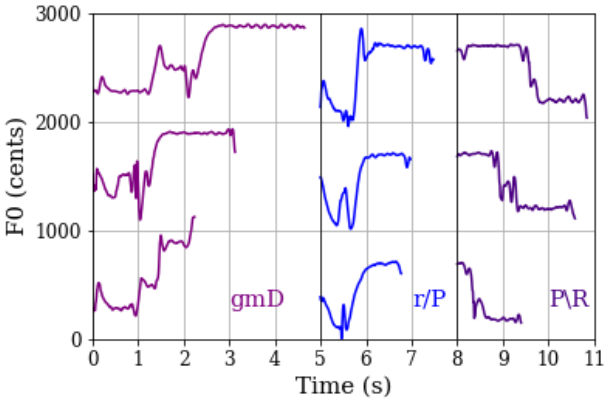

Figure 3: Sample templates for each of the three phrases: gmD (purple), r/P (blue) and P\R (violet). Vertical and horizontal offsets applied for better viewing.

4. EXPERIMENTS

Past work on gesture kinematics in the context of speech and singing co-gesturing has considered velocity and acceleration parameters rather than the raw position time series with these parameters relating more directly to human effort or force [10, 29–31]. We therefore include the (x,y) position of each wrist as well as the corresponding velocity and acceleration profiles across the segment as input features for our two classification tasks. Summarising the previous section, the gesture time series is segmented based on the previously obtained audio segment boundaries giving us a time-aligned multidimensional time series for (i) each stable-note and non-stable segment across all the alaps in the dataset, and (ii) each pakad phrase gesture template and its audio-matched gesture segment from the SDS. We consider supervised classification for each task with the different features as discussed next.

4.1 Stable-note prediction

Stable note segments were labeled as such based on the F0 variation across the segment as discussed in Section 3.1. We would like to investigate whether there is any consistency in the gesture kinematics corresponding to stable note regions that discriminates them from that of the non-stable F0 regions.We implement a binary classifier trained and tested on the dataset of labeled stable notes and the (complementary) non-stable regions where the training and test data are both drawn from across singers and ragas. Although 250 ms regions of stable pitch qualified as stable notes, we restricted the examples of both categories used in this experiment to those with a minimum duration of 1 s in order to ensure that the training dataset was not severely imbalanced. The proportion of stable notes exceeded 20%

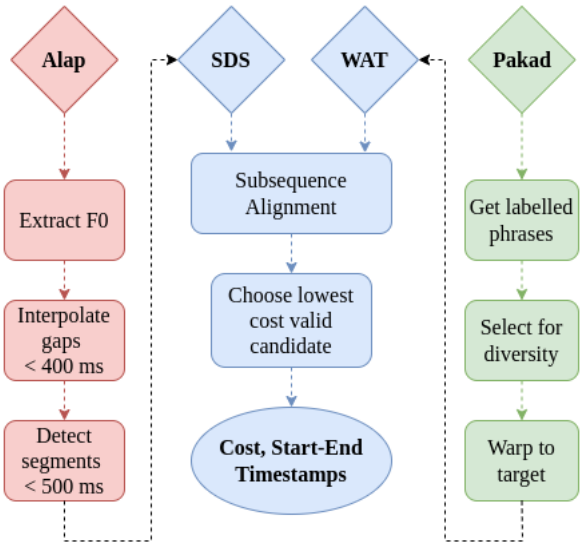

Figure 4: The pipeline for phrase-based segmentation using alap and pakad audio data. For warping the pakad phrase templates, a window size of 100 and a penalty of 200 was chosen, while for the subsequence alignment, K = 20 and penalty = 0.1 were chosen [27].

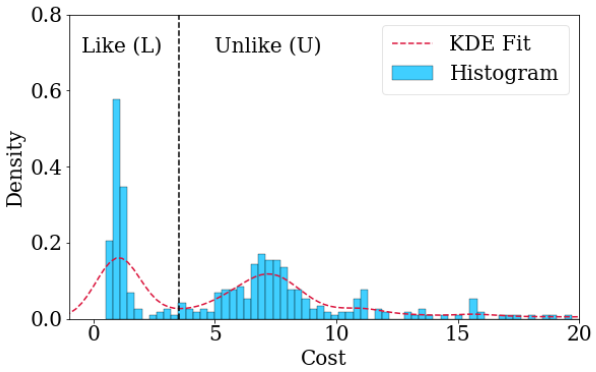

Figure 5: Distribution of the DTW subsequence cost across the SDS of all singer alaps for the best matched audio phrase template for P\R of raga Nand. The dashed vertical line shows the threshold derived from the KDE fit (dashed contour), using which the SDS are labeled as Like and Unlike with reference to the template phrase.

Dur.	Count	Stable (%)	P	P+V	V+A	P+V +A
>1s	34612	24.36	39.6	39.2	**43.8**	40.0
>2s	12000	16.43	26.0	**31.8**	**36.7**	**32.4**

Table 3: F1-score (%) obtained for various combinations of kinematic features of position (P), velocity (V) and acceleration (A). Bold font indicates the model is significantly better than a random baseline ($p<0.001$). The random baseline F1 score for durations >1s is 39.2% and for >2s is 28.2%

of the total examples when the minimum segment duration was constrained to 1 s. Further, segments greater than 5 s in length were truncated to 5 s.

We consider the multiple gesture time series as inputs for our classification between stable-note and non-stable segments. We use a convolutional (1D) classifier after padding / truncating all sequences to 5 second duration (500 timesteps) and the appropriate masking of the padded sequences [32]. We create 4 train-test splits of the data so that every example of the dataset appears in the test set once. The training and validation data consist of 60% and 15% respectively of the full dataset. We train a 1D convolutional neural network on the multidimensional time series and choose the best model architecture using keras tuner [33] with an objective to find the architecture providing the best F1 score on the validation data. Our hyperparameters involve the number of convolutional layers, number of filters, kernel size and number of dense layers and number of nodes in dense layer. We use Adam optimizer with default parameters for training. We evaluate each split of the test data for the chosen trained model and threshold and combine the results across all 4 test splits in Table 3. We repeat this process for various combinations of kinematic variables viz. Position (P), Position and velocity (P+V), velocity and acceleration (V+A) and all (P+V+A).

4.2 Raga phrase detection

Our goal is to determine whether the L and U labels (that were assigned based on audio proximity) can be predicted by gesture alone at better than chance (i.e. based only on the priors) and, if so, which kinematic features are most useful in this task. Our measure of similarity is the DTW distance computed between the template and test (i.e.the alap SDS subsequence) time series. In the context of our alap gesture time series, already segmented based on the audio phrase matching, we now compute DTW distance between the multidimensional reference and candidate under test.

Multidimensional time series present us with some distinct options for the distance computation. Two obvious approaches are DTW_I and DTW_D depending on whether the individual time series are each warped independently or whether they are all forced into a single warping path [34]. The use of DTW_D appears meaningful for the incorporation of the velocity and acceleration contours derived from the corresponding position time series of the wrists. However, it is interesting also to test with independent DTW costs across the separate time series (to get an 8d feature vector of costs) to see if this helps reduce the effect of the less informative features, if any. We term this DTW_{IND}. Further, decoupling the left and right wrists to obtained two differently warped sets of time series (DTW_{LR}) is also perfectly meaningful in the current task.

With DTW cost(s) as the input features, we create 5 train-test splits with the uniform distribution of singers across the splits. Thus every example appears once in the test set. We train a logistic regression classifier with L2 regularizer and use 3-fold cross-validation within the train

Phrase	Like	Unlike	Chance Acc.	DTW_D (1)	DTW_I(1)	DTW_{Ind} (8)	DTW_{LR} (2)
gmD	944	827	50.2	52.2	48.6	51.8	**52.4**
r/P	1035	1268	50.5	**55.3**	47.1	**56.1**	**55.1**
P\R	817	1340	53.0	**65.0**	45.7	**65.2**	**65.1**

Table 4: Classification accuracy (%) for Like and Unlike phrase detection with gesture time series and different DTW distance measures. Feature dimensionality (i.e. DTW path costs) appears in parantheses. Bold font indicates that the model performance is significantly better ($p<0.005$) than chance, with the chance accuracy (%) also mentioned in the table.

set to learn the best set of parameters.

5. RESULTS AND DISCUSSION

We discuss the outcomes of the experiments.

5.1 Stable-note detection

Table 3 presents the stable note detection results in terms of the F1 score for retrieval of the stable notes. As mentioned in Section 4.1, we restrict ourselves to the set of labeled segments of duration greater than 1s, with 33484 examples in all. With 23% of these corresponding to stable notes, we find that the obtained F1 scores across feature sets are slightly above that for a random baseline system (based on the priors only) with none of the differences significant. We therefore consider a subset of longer duration segments (duration > 2 s) for a better opportunity at gesture analysis given the longer segment of the time series now available. We have a reduced count of 12620 examples now, with close to 15% stable notes. As reported in Table 3, we get F1 scores that are significantly better than the chance baseline for all but the raw position time series. The combination of position, velocity and acceleration (P+V+A) performs best.

On examining a breakdown of the misclassifications across singers and ragas (details reported in the supplementary material), we find that the performance is poorer for singers who exhibit a particularly low proportion of stable notes. Such behaviour can arise, for example, when the singer makes a choice to focus more on melodic movements in their alap rather than long periods of held notes. With a very low representation of their stable note examples in the training data, it is probable that idiosyncratic aspects of their stable note gestures, if any, were not learned by the classifier. We did not find much of raga dependence in stable note detection performance. We also did an analysis of tonic versus other stable notes to find that the tonic notes (fewer in number overall) were harder to detect; this observation needs more data for a better understanding.

5.2 Raga-phrase detection

Table 4 displays the Like/Unlike classification of raga phrases across the alaps of all singers. We see a roughly equal proportion of L and U examples and therefore chance baseline accuracies close to 50%. Both P\R and r/P exhibit gesture classification accuracies that are statistically better than chance for all versions of DTW distance except the DTW_I which is the simple summing of independent path costs across the 8 different series. In the case of the gmD phrase, we see a relatively small increase over chance with the only significant difference provided by the DTW_{LR} that combines left and right wrist paths, each computed independently of the other. A singer-based breakdown of the overall accuracy showed relatively uniform behaviour across singers for all the phrases except for one outlier (out of the 11) for each of P\R and r/P phrases.

We would also like to comment on the equal proportion of L and U examples in our data for this task. Although there is a far larger number of U instances (that is alap segments that probably do not contain the phrase of interest and therefore expected to return a high cost in the DTW subsequence search of the audio), we found that many of these actually led to invalid paths from pathological warping and thus were unusable candidates for this study.

6. CONCLUSION

This work proposed a new approach to examining melodic similarity captured in co-singing gestures by analysing audiovisual recordings. With a new dataset of 11 singers, raga-characteristic phrases were proposed as a proxy for similar melodic movements within and across singers. As in previous work, wrist movements that accompanied the solo alap singing were represented as kinematics time series. In the absence of ground-truth phrase labels for the alap data, we developed a pipeline for achieving the AV segmentation for the chosen phrases via DTW-based audio template matching using a small set of hand-labeled segments. We also considered the classification task for more generic AV segments defined in a bottom-up manner such as stable-note regions. Our experimental results indicate classification performances that are marginally above chance even if statistically significant. Further, the importance of including computed velocity and acceleration profiles in the gesture representation has been confirmed.

A useful contribution of this work is the musicological questions it encourages. Apart from the aspects already mentioned in the discussion of the results, we note that the use of multiple phrase templates can facilitate larger experimental validation of hypotheses, such as that of Rahaim [5], that gestures could function to draw attention to *what is different* between two semantically close melodic patterns. Finally, several enhancements to the presented methods are possible including adding more keypoints (elbow and hand joints) and using all 3 camera views to include depth movement.

7. REFERENCES

[1] M. Clayton, "Time, gesture and attention in a khyāl performance," *Asian Music*, vol. 38, no. 2, pp. 71–96, 2007.

[2] L. Leante, "The lotus and the king: Imagery, gesture and meaning in a hindustani rāg," *Ethnomusicology Forum*, vol. 18, no. 2, pp. 185–206, 2009.

[3] Leante, "Gesture and imagery in music performance: Perspectives from north indian classical music," in *The Routledge Companion to Music and Visual Culture*. Routledge, 2013, pp. 145–152.

[4] L. Leante, "The cuckoo's song : imagery and movement in monsoon ragas," in *Monsoon feelings : a history of emotions in the rain*, I. Rajamani, M. Pernau, and K. R. B. Schofield, Eds. New Delhi: Niyogi Books, 2018.

[5] M. Rahaim, *Musicking Bodies: Gesture and Voice in Hindustani Music*. Wesleyan University Press, 2012.

[6] S. Paschalidou, "Effort inference and prediction by acoustic and movement descriptors in interactions with imaginary objects during dhrupad vocal improvisation," *Wearable Technologies*, vol. 3, p. e14, 2022.

[7] M. Clayton, P. Rao, N. Shikarpur, S. Roychowdhury, and J. Li, "Raga classification from vocal performances using multimodal analysis," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR, Bengaluru, India, pp. 283-290.*, 2022.

[8] 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

[9] L. Pearson, "Gesture and the sonic event in karnatak music," *Empirical Musicology Review*, vol. 8, no. 1, pp. 2–14, 2013.

[10] L. Pearson and W. Pouw, "Gesture–vocal coupling in karnatak music performance: A neuro–bodily distributed aesthetic entanglement," *Annals of the New York Academy of Sciences*, vol. 1515, no. 1, pp. 219–236, 2022.

[11] W. Pouw *et al.*, "A kinematic-acoustic analysis of gesture-speech coupling in persons with aphasia," 2021.

[12] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, p. 2154, 2020.

[13] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[14] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor *et al.*, "Essentia: an audio analysis library for music information retrieval," in *Proc. of the 14th Int. Soc. for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[15] 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.

[16] C. Larboulette and S. Gibet, "A review of computable expressive descriptors of human motion," in *Proceedings of the 2nd International Workshop on Movement and Computing*, 2015, pp. 21–28.

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

[18] 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.

[19] W. Pouw, J. de Wit, S. Bögels, M. Rasenberg, B. Milivojevic, and A. Ozyurek, "Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics," in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior: 12th International Conference, DHM 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I*. Springer, 2021, pp. 269–287.

[20] K. Ganguli and P. Rao, "A study of variability in raga motifs in performance contexts," *Journal of New Music Research*, vol. 50, pp. 1–15, 02 2021.

[21] W. Van der Meer, *Hindustani music in the 20th century*. Springer Science & Business Media, 2012.

[22] K. K. Ganguli and P. Rao, "On the distributional representation of ragas: Experiments with allied raga pairs," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.

[23] C. Gupta and P. Rao, "An objective assessment tool for ornamentation in singing," in *Proceedings of the International Symposium of Frontiers of Research on Speech and Music and Computer Music Modelling and Retrieval*, 2011.

[24] "Music in motion, the automatic transcription system for indian music," https://autrimncpa.wordpress.com/, note = Last Accessed: 2023-04-14.

[25] S. Kulkarni, *Shyamrao Gharana*. Prism Books Pvt. Ltd, 2017, vol. 1.

[26] M. Müller, *Fundamentals of Music Processing*. Springer, 2015.

[27] T. V. C. . P. R. Wannes Meert, Kilian Hendrickx, "Dtaidistance (version v2)," last Accessed: 2023-04-14. [Online]. Available: http://doi.org/10.5281/zenodo.5901139

[28] S.-T. Chiu, "Bandwidth selection for kernel density estimation," *The Annals of Statistics*, pp. 1883–1905, 1991.

[29] R. C. Madeo, C. A. Lima, and S. M. Peres, "Gesture unit segmentation using support vector machines: segmenting gestures from rest positions," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 46–52.

[30] W. Pouw and J. A. Dixon, "Quantifying gesture-speech synchrony," in *the 6th gesture and speech in interaction conference*. Universitaetsbibliothek Paderborn, 2019, pp. 75–80.

[31] Y. Ferstl, M. Neff, and R. McDonnell, "Expressgesture: Expressive gesture generation from speech through database matching," *Computer Animation and Virtual Worlds*, vol. 32, no. 3-4, p. e2016, 2021.

[32] "Understanding masking & padding," https://keras.io/guides/understanding_masking_and_padding/, (Accessed on 04/15/2023).

[33] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "Kerastuner," https://github.com/keras-team/keras-tuner, 2019.

[34] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach," *Data mining and knowledge discovery*, vol. 31, pp. 1–31, 2017.