A SCORE-INFORMED COMPUTATIONAL DESCRIPTION OF SVARAS USING A STATISTICAL MODEL

Sertan Şentürk, Gopala Krishna Koduri, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{sertan.senturk, gopala.koduri, xavier.serra}@upf.edu

ABSTRACT

Musical notes are often modeled as a discrete sequence of points on a frequency spectrum with possibly different interval sizes such as just-intonation. Computational descriptions abstracting the pitch content in audio music recordings have used this model, with reasonable success in several information retrieval tasks. In this paper, we argue that this model restricts a deeper understanding of the pitch content. First, we discuss a statistical model of musical notes which widens the scope of the current one and opens up possibilities to create new ways to describe the pitch content. Then we present a computational approach that partially aligns the audio recording with its music score in a hierarchical manner first at metrical cycle-level and then at note-level, to describe the pitch content using this model. It is evaluated extrinsically in a classification test using a public dataset and the result is shown to be significantly better compared to a state-of-the-art approach. Further, similar results obtained on a more challenging dataset which we have put together, reinforces that our approach outperforms the other.

1. INTRODUCTION

A musical note can be defined as a sound with a definite pitch and a given duration. An interval is a difference between any two given pitches. Most melodic music traditions can be characterized with a set of notes it uses and the corresponding intervals. They constitute the core subject matter of research concerning the tonality and melodies of a music system. For any quantitative analyses therein, it is required to have a working definition and a consequent computational model of notes which dictate how and what we understand of the pitch content in a music recording.

In much of the research in music analysis and information retrieval, the most commonly encountered model is one that considers notes as a sequence of points separated by certain intervals on frequency spectrum. There are different representations of the pitch content from a given recording based on this notion, the choice among which is influenced to a great degree by the intended application. Examples include pitch class profiles [1], harmonic pitch class profiles [2], pitch histograms [3] and pitch kernel density estimates [4] besides others.

Albeit a useful model of notes used alongside several information retrieval tasks, we believe it is limited in its purview. To elaborate, we consider the case of Carnatic music, an art music tradition from south India. The counterpart to note in this tradition is referred to as svara, which has a very different musicological formulation. A svara is defined to be a definite pitch value with a range of variability around it owing to the characteristic movements arising from its melodic context. The seven svaras in Carnatic music are $S(a), R(i), G(a), M(a), P(a), D(ha), N(i)$, which account for 12 pitch positions (svarasthanas), S, R_1, R_2/G_1, R_3/G_2, G_3, M_1, M_2, P, D_1, D_2/N_1, D_3/N_2, N_3 [5]. It is emphasized that the identity of a svara lies in this variability [5], which makes it evident that the former model of notes has a very limited use in this case. The arguments related to variability are also relevant to Hindustani music, an art music form prevalent in northern parts of the Indian subcontinent and as well as many other melody-dominant music cultures such as Ottoman-Turkish makam music.

In this paper, we discuss a statistical model of notes that broadens the scope of the former, encapsulating the notion of the variability in svaras (Section 3). We develop a methodology that exploits score information to automatically process the pitch content of audio recordings (Section 4). The methodology first aligns the audio recording with the relevant music score. This step is designed to handle the structural differences between the music score and the audio performance. Next, the pitch values are aggregated for each note symbol from the aligned instances of the notes and these pitch values are used to compute a statistical representation for each note. The methodology is evaluated extrinsically in a classification task comparing the results with a state-of-the-art system [6] (Section 5) using two datasets (Section 2).

Our contributions in this paper can be summarized as:

1. A novel, computational note model, which is able to describe the characteristics of the notes statistically besides its definite location
2. Adaptation of a state of the art audio-score alignment method proposed for another melody dominant culture to Carnatic music
3. Simplifications and generalizations on the adapted audio-score alignment method
4. A new dataset of Carnatic music, composed of audio recordings and music scores linked to each other in the document-level

Raaga	#Comp.	#Singer	#Rec.
Anandabhairavi	3	5	7
Atana	4	5	5
Bhairavi	5	7	8
Devagandhari	5	5	5
Kalyani	4	4	5
Todi	9	15	15
Total	**30**	**24**	**45**

Table 1. A more diverse dataset compared to the Carnatic Varnam dataset. This consists of 40 recordings in 6 raagas performed by 24 unique singers encompassing 30 compositions.

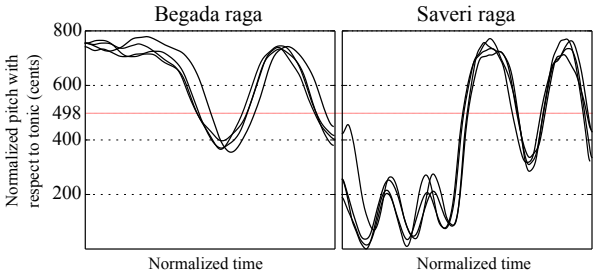

Figure 1. Example pitch contours of M_1 svara in different raagas. the X-axis is time normalized with respect to the length of each pitch contour. The tuning of M_1 svara according to the just-intomation temperament (498 cents) is indicated with a continuous red line. Notice that the majority of the pitches are sung quite distant from the theoretical tuning.

2. DATA

For evaluation, we use the Carnatic Varnam dataset [1] (see [6] for a description of varnams and the dataset). Varnams are compositions that are often sung to the score unlike several other forms which are interlaced with improvisation. Note that even though the order of the cycles in the score are retained, the performers tend to omit a few cycles or repeat a few of them twice with some minor variations. The dataset has annotations at the metrical cycle-level synchronizing the audio recording and the extracted melody with the score. There are 7 raagas, 27 recordings and 1155 cycle-level annotations. The average cycle-duration is 9.8 seconds with a standard deviation of 1.2 seconds. The music scores in the dataset are notated as a sequence of svara symbols and their relative durations. The metrical cycles are indicated in the score. There is no nominal tempo information in the score as the performance tempo is decided by the performer. With an assumption that each svara within the cycle is sung exactly according to its relative duration in the score, the svaras in the recording are annotated semi-automatically.

This dataset comes with a limitation that all the performances of a given raaga are of the same composition. Therefore the representation computed for a svara can be specific to either the raaga or the composition. In order to eliminate this ambiguity, we have put together another dataset, which is more diverse in terms of the number of compositions per raaga. [2] The details of the dataset are shared in Table 1. The Carnatic Varnam dataset is drawn from the performances of a compositional form known as varnam. Our dataset contains performances of another compositional form known as kriti. The latter are more common in concert performances, where the performers take liberty to do an impromptu improvisation. As a result, kritis are almost always not sung to the score and hence pose more challenges compared to varnams for a score-informed approach such as ours. Note that we follow the same format of the scores in Carnatic Varnam dataset to notate the kriti compositions.

[1] Available at http://compmusic.upf.edu/carnatic-varnam-dataset

[2] The dataset is available at `http://compmusic.upf.edu/node/314`.

3. MODEL OF MUSICAL NOTES

Research that involved analysis of svaras in Indian art music has time and again shown that reducing svara to a frequency value results in loss of important information [4, 7, 8]. Computational svara descriptions that use more melodic context for the description of a svara such as pitch histograms, have been shown to outperform the naive descriptions such as pitch-class distributions [6, 9]. We build on these observations from the past research and consolidate that to a statistical model of notes that would facilitate extracting information that is otherwise opaque to the currently used model.

Figure 1 shows melodic contours extracted from the individual recordings of M_1 svara (498 cents in just-intonation) in different raagas. It shows that a svara is a continuum of varying pitches of different durations, and the same svara is sung differently in two given raagas. Note that a svara can vary even within a raaga in its different contexts [7,8]. Taking this into consideration, we propose a statistical model of notes that aims for a more inclusive representation of pitches constituent in a svara. In this model, we define a note as *a probabilistic phenomenon on a frequency spectrum*. This notion can be explored in two approaches that are complementary in nature: i) *temporal*, which helps to understand the evolution of a particular instance of a svara over time (This has been theoretically explored in [8]) and ii) *aggregative*, which allows for studying the whole pitch space of a given svara in its various forms, often discarding the time information.

Our method, presented in the following section, takes the latter approach. From the annotations in our dataset, we aggregate the pitch contours over the svara reported in Figure 1 for the same set of raagas. Figure 2 shows its representations, computed as described in Section 4.2. The correspondences between the two figures are quite evident. For instance, M_1 in Begada is sung as an oscillation between G_3 (386 cents) and M_1. The representation reflects this with peaks at the corresponding places. Further, the shape of the distributions reflect the nature of pitch activity therein. The goal of our approach is to obtain such representations for svaras across different raagas in our dataset

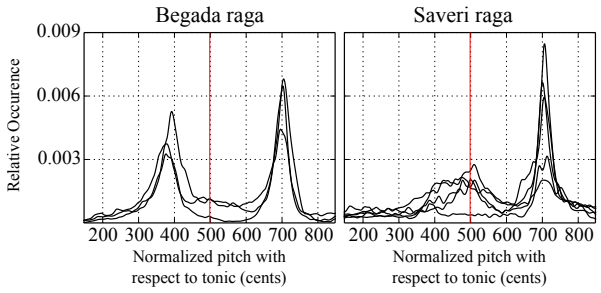

Figure 2. Histograms of M_1 svara computed from the annotated pitch contours shown in Figure 1. The tuning of M_1 svara according to the just-intomation temperament (498 cents) is indicated with continuous red lines.

automatically.

4. METHODOLOGY

Our method starts by aligning the audio and the score at the cycle- and the svara-level (Section 4.1). Then the pitch values in different instances of a given svara are obtained and an aggregate representation of a svara is computed (Section 4.2).

4.1 Audio-score alignment

Audio-score alignment can be defined as *the process of finding the segments in the audio recording that correspond to the performance of each musical element in the music score*. For this task, several approaches have been proposed using techniques such as Hidden Markov models [10, 11], conditional random fields [12] and dynamic time warping [13–15].

The structural mismatch between the music score and the audio recording is a typically encountered challenge in audio-score alignment. This is also common phenomenon in the performances of varnams and kritis, where the singers tend to repeat, omit or insert cycles in the score. To overcome this problem there exists methodologies, which allow jumps between structural elements [14, 16]. However these methodologies are not designed to skip musical events in the performance, which are not indicated in the score, such as impromptu improvisations commonly sung in kritis (Section 2). Moreover, we may not need a complete alignment between the score and audio recording in order to accumulate a sufficient number of samples for each svara.

In [17], an audio-score alignment methodology for aligning audio recordings of Ottoman-Turkish makam music with structural differences and events unrelated to the music score was introduced, and it is later extended to note-level alignment in [18]. The methodology proposed in [17], divides the score into meaningful structural elements using the editoral section annotations in the score. It extracts a predominant melody from the audio recording and computes a synthetic pitch of each structural element in the score. Then it computes a binarized similarity matrix for each structural element in the score from the predominant melody extracted from the audio recording and the synthetic pitch. The similarity matrix has blobs resembling lines positioned diagonally, indicating candidate alignment paths between the audio and the structural element in the score. Hough transform, a simple and robust line detection method [19], is used to locate these blobs and candidate time-intervals for where the structural element is performed is estimated. To eliminate erroneous estimations, [17] uses a variable-length Markov model based scheme, which is trained on structure sequences labeled in annotated recordings. Finally, Subsequence Dynamic Time Warping (SDTW) is applied to the remaining structural alignments to obtain the note-level alignment [18].

Our alignment methodology is based on the procedure described in [17, 18]. Since the original methodology is proposed for Ottoman-Turkish makam music, we optimize several parameters according to the characteristics of our data. We also modify several steps in the original methodology for the sake of generalization and simplicity. These changes will be detailed throughout this section, hereafter. The procedure in our methodology can be summarized as:

1. Extract features from the audio recording and the music score (Section 4.1.1)
2. Estimate possible partial alignments between the audio recording and the score in the cycle-level (Section 4.1.2)
3. Discard erroneous estimations (Section 4.1.3)
4. Extract svara samples from the note-level alignment within each aligned cycle. (Section 4.1.4)

4.1.1 Feature Extraction

Given an audio recording, we extract a predominant melody using the method proposed in [20], which has been shown to output reliable pitch estimations on Carnatic music recordings [6]. We denote the predominant melody extracted from the audio recording as $f = \big(f_1, \ldots, f_V\big)$, where V is the number of samples in the predominant melody. The sampling rate of the predominant melody is equal to ≈ 334.5 Hz, which is reported as an optimal for the methodology in [20]. Note that the timestamp of a pitch sample, f_i, is denoted as $\tau(f_i)$.

We then normalize the pitch values, $f_i \in f$, from Hz to cent scale with respect to the tonic frequency, t, by:

$$x_i = 1200 \log_2\Big(\frac{f_i}{t}\Big) \tag{1}$$

Note that there are 1200 cents in an octave. The tonic is extracted automatically using [21], which is reported to output near-perfect results in identifying the tonic of Carnatic music recordings. We denote the normalized predominant melody extracted from the audio recording as $x = \big(x_1, \ldots, x_V\big)$.

Parallel to audio predominant melody extraction, the svara symbols notated in the score are mapped to their cent-scale equivalents using just-intonation temperament [22]. Then, the score is divided into cycles according to the cycle boundaries annotated in the score. For each cycle (n), a synthetic pitch is computed by sampling a hypothetical continuous pitch contour corresponding to the svara sequence [17]. In this process, we consider the tempo of

the score as 70 bpm, which is reported in [9] as the average tempo in the Carnatic Varnam dataset. We denote the synthetic pitch of cycle (n) as $y^{(n)} = \big(y_1^{(n)}, \ldots, y_{W^{(n)}}^{(n)}\big)$, $n \in [1:N]$, where N is the number of cycles in the score and $W^{(n)}$ is the number of samples in the synthetic pitch. The sampling rate of the synthetic pitch is equal to the sampling frequency of the audio predominant melody. During the synthetic pitch computation, the svara onset and offset timestamps are recorded. We will use this information to obtain the svara-level alignment (Section 4.1.4) later.

4.1.2 Estimating cycle-level alignment

Instead of Hough transform used in [17], we use Iterative Subsequence Dynamic Time Warping (ISDTW) [23, Chapter 4], a common methodology used to find a queried subsequence in a given target [24, 25] to estimate the time-intervals, where a cycle is performed. Our preliminary experiments on the Carnatic Varnam Dataset showed that using ISDTW gave comparable results to Hough transform. Moreover, ISDTW simplifies the note-level alignment step compared to [18] since note onset and offsets can be directly inferred from the paths obtained from ISDTW, without introducing an additional process (e.g. SDTW in [18]) as described in Section 4.1.4.

We set the step size to $\{(2,1),(1,1),(1,2)\}$. This step size restricts the path between half and double of the tempo, which helps to avoid pathological errors. To obtain an accumulated cost matrix, $C^{(n)}$ for each cycle (n), we use the local distance measure:

$$d\left(x_i, y_j^{(n)}\right) = min\Big(\big(|x_i - y_j^{(n)}| \bmod 1200\big), \\ 1200 - \big(|x_i - y_j^{(n)}| \bmod 1200\big)\Big) \quad (2)$$

where x_i and $y_j^{(n)}$ denote the i and j^{th} samples of the audio predominant pitch x and synthetic pitch $y^{(n)}$, respectively. This distance may be interpreted as the shortest distance in cents between two pitch classes. It is not affected by octave-errors in the normalized predominant melody [17].

We use the iterative algorithm given in [23, Page 81] to estimate multiple alignments for each cycle (n). We iterate the algorithm for 10 times for each cycle. After each iteration, we obtain an estimation $e^{(k,n)}$ with an optimal alignment, $p^{(k,n)} = \big(p_1^{(k,n)} \ldots p_{L^{(k,n)}}^{(k,n)}\big)$ with $p_l^{(k,n)} = \big(r_l^{(k,n)}, q_l^{(k,n)}\big)$, $r_l^{(k,n)} \in x$, $q_l^{(k,n)} \in y^{(n)}$, $l \in \big[1 : L^{(k,n)}\big]$ (where $L^{(k,n)}$ is the length of the alignment $p^{(k,n)}$) and $k \in [1:10]$ (since there are 10 iterations for each cycle). The estimated time-interval, $t^{(k,n)}$, is the subsequence of the audio recording in the time-interval $[\tau\big(r_1^{(k,n)}\big) : \quad \tau\big(r_{L^{(k,n)}}^{(k,n)}\big)]$. For each alignment we also record the cost at each step as:

$$d^{(k,n)} = \big(d_1^{(k,n)}, \ldots, d_{L^{(k,n)}}^{(k,n)}\big) \\ = \Big(d\big(r_1^{(k,n)}, q_1^{(k,n)}\big), \ldots, d\big(r_{L^{(k,n)}}^{(k,n)}, q_{L^{(k,n)}}^{(k,n)}\big)\Big) \quad (3)$$

After each iteration, we set the values between $r_l^{(k,n)} \pm 0.1W^{(n)}$ in the accumulated cost matrix, $C^{(n)}$, to infinity for the next iteration to ensure a new path will not be searched nearby. Remember $W^{(n)}$ is the number of samples in the synthetic pitch, $y^{(n)}$.

To distinguish correct alignments from the erroneous, we compute a similarity value $s^{(k,n)} \in [0:1]$ for an iteration (k) of the cycle (n). We use the similarity measure between the cycle and the estimated alignment proposed by [17, Page 15, described as weight normalization]:

$$s^{(k,n)} = \frac{\sum_l^{L^{(k,n)}} \beta(p_l^{(k,n)}, q_l^{(k,n)})}{L^{(k,n)}} \quad (4)$$

where the binarization criteria is defined as:

$$\beta(a,b) = \begin{cases} 1, & d\big(a,b\big) \leq \alpha \\ 0, & d\big(a,b\big) > \alpha \end{cases} \quad (5)$$

In Section 5, we present the experiments to find the optimal value for the binarization threshold, α. The true positives are observed to typically emit a higher score than the erroneous ones. Performing the ISDTW for each cycle, we obtain estimations $e^{(k,n)} = \big\{n, t^{(k,n)}, p^{(k,n)}, s^{(k,n)}\big\}$, where n is the cycle extracted from the score, $t^{(k,n)}$ is the estimated time-interval in the audio recoding, $p^{(k,n)}$ is the optimal alignment of the estimation and $s^{(k,n)}$ is the similarity value of the estimation.

4.1.3 Discarding erroneous estimations

At this step we obtain a considerable number of correct estimations albeit with a comparable number of erroneous estimations. Nonetheless, we need to ensure a high precision in the cycle-level alignment to obtain a reliable svara description. In order to achieve this we can afford to trade the recall in the process since a moderate recall in the cycle-level alignment would still be able to supply a good number of samples per svara.

The method proposed for discarding erroneous estimations in [17] is not generalizable as introducing a new form with a different structure requires substantial number of training recordings in that form. For this reason, we choose to use an unsupervised estimation selection scheme, which is more generalizable and simpler.

We classify the estimations into two classes with respect to their similarity values using k-means clustering [26]. We use squared Euclidean distance as the distance measure and discard the cluster with low scores. Next, we check if there are estimations, which overlap more than 3 seconds in time. In such a case we only keep the estimation with the highest similarity value as the music has a single melody track throughout. In Section 5, we report alignment results after discarding estimations both without (i.e. only discarding overlapping estimates) and with k-means clustering.

4.1.4 Svara-level alignment

Recall that the svara onset and offset timestamps in each cycle of the synthetic pitch, $y^{(n)}$, are known. The aligned svara onset and offsets are directly obtained as the timestamps $\tau\big(r_l^{(k,n)}\big)$, which are mapped to these onsets and offsets inside the alignment $p^{(k,n)} = \big(p_1^{(k,n)} \ldots p_{L^{(k,n)}}^{(k,n)}\big)$, respectively.

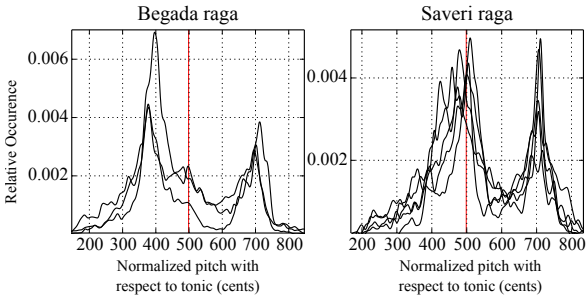

Figure 3. Description of M_1 svara (498 cents in just intonation) using our approach.

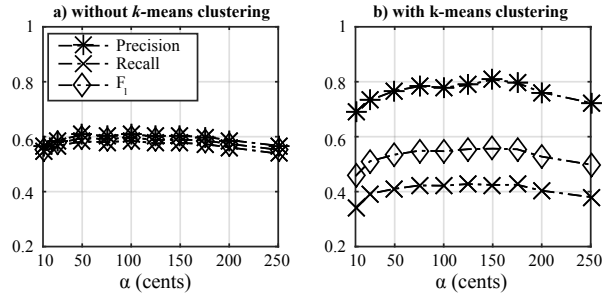

Figure 4. Results of cycle-level alignment for different binarization threshold values.

4.2 Computing svara representations

For a given recording, for each svara, σ, in the corresponding raaga, we obtain a pool of normalized pitch values, $x^\sigma = \{x_1^\sigma, x_2^\sigma, \dots\}$, aggregated over all the aligned instances from its melodic contour (Section 4.1.4). Our representation must capture the probabilities of the pitch values in a given svara. Histograms are a convenient way for representing the probability density estimates [4,6]. Therefore, we compute a normalized histogram over the pool of the pitch values. For brevity sake, we consider pitch values over the middle octave (i.e., starting from the tonic) at a bin-resolution of one cent:

$$h_m^\sigma = \frac{\sum_i \lambda_m(x_i^\sigma)}{|x^\sigma|}, \tag{6}$$

where h_m^σ is the probability estimate of the m-th bin, $|x^\sigma|$ is the number of pitch values in x^σ and λ function is defined as:

$$\lambda_m(a) = \begin{cases} 1, & c_m \leq a \leq c_{m+1} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where a is a normalized pitch sample and (c_m, c_{m+1}) are the bounds of the m-th bin.

Figure 3 shows the representations obtained in this manner for M_1 svara (our running example from Figure 1) in different raagas. Notice that the representations obtained for M_1 are similar to the corresponding representations shown in Figure 2. This representation allows to deduce important characteristics of a svara besides its definite location (i.e., 498 cents) in the frequency spectrum. For instance, from Figure 3, one can infer that M_1 in Begada and Saveri are sung with an oscillation that ranges from G_3 (386 cents) to P (701 cents) in the former and M_1 to P in the latter.

5. EVALUATION AND RESULTS

Our method is evaluated on the two datasets described in Section 1 using the following tasks:

i. The cycle-level alignment, evaluated intrinsically using the ground truth annotations from the Carnatic Varnam dataset.

ii. The svara-level alignment and the computed representation, evaluated extrinsically using a raaga classification task on both the datasets.

The svara-level alignments cannot be verified in an intrinsic manner because marking the ground truth is prone to be erroneous as it is difficult for even musicians to agree with each other on the exact boundaries of a svara sung in a melodic continuum. [3]

To evaluate the cycle-level alignment, we check the time-distance between the estimated borders of the cycle and annotated borders as described in [17]. A cycle is marked as a true positive if the distance between both of the boundaries of the aligned cycle and the relevant annotation is less than 3 seconds. It is marked as a false positive otherwise. If there is no estimation for an annotation, it is marked as a false negative.

Figure 4 shows the recall, precision and F_1-score for different binarization thresholds used in similarity computation. Figure 4a shows that our methodology achieves a balanced recall and precision in the cycle-level alignment even without having a precise information on the performance tempo. Figure 4b shows that the process described to the discard erroneous alignments (Section 4.1.3) removes most of the false positives within an acceptable decrease in recall. It can also be observed that our cycle-level alignment is insensitive to the binarization threshold, α. When the parameter is selected between 50 cents (a quarter tone) and 200 cents (a whole tone), there is no a significant difference in the alignment results at the $p = 0.01$ level as determined by a multiple comparison test using the Tukey-Kramer statistic. Hereafter, we report results for a binarization threshold of 150 cents.

Using an α of 150 cents, we achieve a 0.42 recall, 0.81 precision and 0.56 F_1-score in cycle-level alignment after discarding the erroneous estimations. The mean and the standard deviation of the true positives are 0.62 and 0.59 seconds, respectively. Within the Carnatic Varnam dataset, we align 606 cycles and 15795 svaras in total. Out of these cycles 490 are true positives. By inspecting the false positives we observed two interesting cases: occasionally an estimated cycle is marked as false positive when one of the boundary distances is slightly more than 3 seconds. The second case is when the melody of

[3] The experiments and the results are available at `http://compmusic.upf.edu/node/314`.

Method	Carnatic Varnam dataset	Our dataset (Table. 1)
Context-based svara distributions [6]	0.62	0.64
Our approach	0.95 to 1	0.88
Using the groundtruth annotations	0.95	N/A

Table 2. Results of raaga classification task over the two datasets using different approaches.

the aligned cycle and performance is similar to each other ($s^{(k,n)} > 0.6$). In both situations considerable number of the note-level alignments would still be useful for the svara model. Within our kriti dataset, 1938 cycles and 59209 svaras are aligned in total.

We use a raaga classification task to evaluate the correctness of the svara alignments and the usefulness of the svara representation created using our statistical model. Our svara representation was shown to perform better compared to the existing representations in our previous work [6]. Therefore, in this task our primary motive is to evaluate the correctness of the svara alignments. However, as marking the svara boundaries is not a viable task, we combine it with evaluating the usefulness of the representation itself in a raaga classification task. We parametrize the representation of each svara using a set of features proposed in our aforementioned work, which include salient observations and the shape parameters of the histogram:

i. The highest probability value in the histogram of the svara
ii. The pitch value corresponding to the highest probability
iii. A probability-weighted mean of pitch values
iv. Pearson's second skewness coefficient
v. Fisher's kurtosis
vi. Variance

There are 12 svaras in Carnatic music, where each raaga has a subset of them. For the svaras absent in a given raaga, we set the features to a nonsensical value. Each recording therefore has 72 features in total.

The smallest raaga-class has three recordings in the Carnatic Varnam dataset, with few classes having more, so we subsampled the dataset six times (corresponding to the highest number of recordings for a class) with three recordings per class. We have also subsampled our dataset in a similar manner. The k-nearest neighbors classifier was earlier shown to perform the best in several raaga classification tasks with varied feature sets [6]. We use the same, with Euclidean distance metric and the number of neighbors set to one.

We compare the results of our approach with the one proposed by Koduri *et al.* [6] which was shown to outperform the previous methods of raaga classification by a slight margin. Their approach uses a moving window to estimate the local temporal context of a small section of melodic contour which is further used to estimate the svara sung at that instance. For each svara, we obtain the corresponding pitch values and use them to create a representation using the method described in Section 4.2, and parametrize it as described earlier in this section. We further compare these results with that obtained using the representation computed from the annotated svara instances in the dataset.

We performed the classification experiment over the subsamples of the two datasets using the leave-one-out cross-validation technique. For our approach, we repeated the experiment with the alignment data resulting from different binarization thresholds. The mean F_1-scores using the representations obtained from the annotations in the dataset, our approach and [6] across the subsampled datasets for the two datasets are reported in Table. 2. Our approach has performed significantly better than the earlier one in [6] on both datasets, and is on par with the method using annotated data. This is a strong indication that our description using the statistical model succeeds in capturing the variability, and therefore the identity of svaras. We also observed that different binarization threshold values have a unimportant impact on the classification accuracy.

6. CONCLUSIONS

We have presented a statistical model of musical notes that expands the scope of the current model in use by addressing the notion of variability of svaras. An approach that builds on this model and exploits scores to describe pitch content in the audio music recordings is presented and evaluated at various levels. The results clearly indicate that our approach is successful in obtaining a computational description of the svaras improving over the state-of-the-art results significantly.

The Carnatic Varnam dataset has 7 raagas, one composition per raaga sung by 3 to 5 artists. We believe this to be one of the contributing factors to a near perfect result using our approach in the raaga classification test. We have put together a more diverse dataset that encompasses more compositions per raaga. Our approach has been shown to be robust to the variability of svaras across compositions in a given raaga. However, we seek attention to the fact that our alignment method relies on the average tempo of the recordings computed from the annotations of the Carnatic Varnam dataset [6]. In order to make the system more self-reliant, we plan to add an initial tempo estimation step similar to [16] by aligning a single cycle using SDTW and resynthesizing the synthetic pitches according to the estimated tempo. We also plan to improve the alignment step by incorporating the svara models in the similarity computation within a feedback mechanism.

An interesting direction to our work is to infer possible facts about a svara from its description. For instance, answering questions such as: i) Is the svara sung steadily? ii) Where is the oscillation on a svara anchored? and so on. These can further be used as parameters that describe

the svara even more concisely. Another direction which interests us is the development of alternative computational descriptions using our statistical model of notes.

Acknowledgments

This research was partly funded by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

7. REFERENCES

[1] T. Fujishima, "Realtime chord recognition of musical sound: A system using Common Lisp Music," in *International Computer Music Conference*, 1999, pp. 464–467.

[2] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, 2006.

[3] A. Gedik and B. Bozkurt, "Pitch-frequency histogram-based music information retrieval for Turkish music," *Signal Processing*, vol. 90, no. 4, pp. 1049–1063, Apr. 2010.

[4] P. Chordia and S. Şentürk, "Joint recognition of raag and tonic in North Indian music," *Journal of New Music Research*, vol. 37, no. 3, pp. 82–98, 2013.

[5] T. M. Krishna and V. Ishwar, "Karṇāṭik music: Svara, gamaka, phraseology and rāga identity," in *2nd CompMusic Workshop*, 2012, pp. 12–18.

[6] G. K. Koduri, V. Ishwar, J. Serrà, and X. Serra, "Intonation analysis of rāgas in Carnatic music," *Journal of New Music Research*, vol. 43, no. 01, pp. 72–93, Jan. 2014.

[7] M. Subramanian, "Carnatic ragam thodi – pitch analysis of notes and gamakams," *Journal of the Sangeet Natak Akademi*, vol. XLI, no. 1, pp. 3–28, 2007.

[8] A. Krishnaswamy, "On the twelve basic intervals in south Indian classical music," *Audio Engineering Society Convention*, pp. 1–14, 2003.

[9] G. K. Koduri, S. Gulati, P. Rao, and X. Serra, "Rāga recognition based on pitch distribution methods," *Journal of New Music Research*, vol. 41, no. 4, pp. 337–350, 2012.

[10] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 974–987, 2010.

[11] A. Maezawa, H. G. Okuno, T. Ogata, and M. Goto, "Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 185–188.

[12] C. Joder, S. Essid, and S. Member, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2385–2397, 2010.

[13] S. Dixon and G. Widmer, "Match: A music alignment tool chest," in *International Society for Music Information Retrieval Conference*, 2005, pp. 492–497.

[14] C. Fremerey, M. Müller, and M. Clausen, "Handling repeats and jumps in score-performance synchronization," in *International Society for Music Information Retrieval Conference*, 2010, pp. 243–248.

[15] B. Niedermayer, "Accurate audio-to-score alignment – data acquisition in the context of computational musicology," Ph.D. dissertation, Johannes Kepler Universität, 2012.

[16] A. Holzapfel, U. Şimşekli, S. Şentürk, and A. T. Cemgil, "Section-level modeling of musical audio for linking performances to scores in Turkish makam music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia: IEEE, 2015, pp. 141–145.

[17] S. Şentürk, A. Holzapfel, and X. Serra, "Linking scores and audio recordings in makam music of Turkey," *Journal of New Music Research*, vol. 43, no. 1, pp. 34–52, 3 2014.

[18] S. Şentürk, S. Gulati, and X. Serra, "Towards alignment of score and audio recordings of Ottoman-Turkish makam music," in *International Workshop on Folk Music Analysis*. Istanbul, Turkey: Computer Engineering Department, Boğaziçi University, 2014, pp. 57–60.

[19] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[20] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, Aug. 2012.

[21] S. Gulati, "A tonic identification approach for Indian art music," Masters Thesis, Universitat Pompeu Fabra, 2012.

[22] J. Serrà, G. K. Koduri, M. Miron, and X. Serra, "Assessing the tuning of sung Indian classical music," in *International Society for Music Information Retrieval Conference*, 2011, pp. 263–268.

[23] M. Müller, *Information retrieval for music and motion*. Springer, 2007.

[24] M. Müller and D. Appelt, "Path-constrained partial music synchronization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 65–68.

[25] X. Anguera and M. Ferrarons, "Memory efficient subsequence DTW for Query-by-Example spoken term detection," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2013, pp. 1–6.

[26] D. J. C. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.