Landmark Detection in Hindustani Music Melodies

Sankalp Gulati[1]
sankalp.gulati@upf.edu

Joan Serrà[2]
jserra@iiia.csic.es

Kaustuv K. Ganguli[3]
kaustuvkanti@ee.iitb.ac.in

Xavier Serra[1]
xavier.serra@upf.edu

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain
[3] DAP Lab, Indian Institute of Technology Bombay, Mumbai, India

ABSTRACT

Musical melodies contain hierarchically organized events, where some events are more salient than others, acting as melodic landmarks. In Hindustani music melodies, an important landmark is the occurrence of a nyās. Occurrence of nyās is crucial to build and sustain the format of a rāg and mark the boundaries of melodic motifs. Detection of nyās segments is relevant to tasks such as melody segmentation, motif discovery and rāg recognition. However, detection of nyās segments is challenging as these segments do not follow explicit set of rules in terms of segment length, contour characteristics, and melodic context. In this paper we propose a method for the automatic detection of nyās segments in Hindustani music melodies. It consists of two main steps: a segmentation step that incorporates domain knowledge in order to facilitate the placement of nyās boundaries, and a segment classification step that is based on a series of musically motivated pitch contour features. The proposed method obtains significant accuracies for a heterogeneous data set of 20 audio music recordings containing 1257 nyās svar occurrences and total duration of 1.5 hours. Further, we show that the proposed segmentation strategy significantly improves over a classical piece-wise linear segmentation approach.

1. INTRODUCTION

Musical melodies contain hierarchically organized events that follow a specific grammar [1]. Some of these events are musically more salient than others and act as melodic landmarks. Cadential notes in classical Western music [2] or Kārvai regions in Carnatic music [3] are examples of such landmarks. While some of these landmarks can be identified based on a fixed set of rules, others do not follow any explicit set of rules and are learned implicitly by a musician through music education and practice. A computational analysis of these landmarks can discover some of these implicitly learned rules and help in developing musically aware tools for music exploration, understanding and education. Occurrence of a nyās in Hindustani music melodies is an example of such a melodic landmark that we investigate in this study.

Dey presents various interpretations and perspectives on the concept of nyās in Hindustani music according to ancient, medieval and modern authors [4]. In the context of its current form, the author describes nyās as that process in a performance of a rāg where an artist pauses on a particular svar [1], in order to build and subsequently sustain the format of a rāg, the melodic framework in Indian art music [4, p. 70][5]. Dey elaborates the concept of nyās in terms of action, subject, medium, purpose and effect associated with it. Typically, occurrence of a nyās delimits melodic phrases (motifs), which constitute one of the most important characteristic of a rāg. Analysis of nyās is thus a crucial step towards melodic analysis of Hindustani music. In particular, automatically detecting occurrences of nyās (from now on referred as nyās segments) will aid in computational analyses such as melody segmentation, motif discovery, rāg recognition and music transcription [6, 7]. However, detection of nyās segments is a challenging computational task, as the prescriptive definition of nyās is very broad, and there are no fixed set of explicit rules to quantify this concept [4, p. 73]. It is through rigorous practice that a seasoned artist acquires perfection in the usage of nyās, complying to the rāg grammar and exploring creativity through improvisation at the same time.

From a computational perspective, the detection of nyās segments is challenging due to the variability in segment length, melodic characteristics and the different melodic contexts in which nyās is rendered. To illustrate this point we show a fragment of pitch contour in Figure 1, annotated with nyās segments denoted by N_i ($i = 1...5$). We see that the nyās segment length is highly varied, where N_5 is the smallest nyās segment (even smaller than many non-nyās segments) and N_3 is the longest nyās segment. In addition, pitch contour characteristics also vary a lot due to the

[1] The seven solfège symbols used in Indian art music are termed as svars. It is analogous to note in western music but conceptually different.

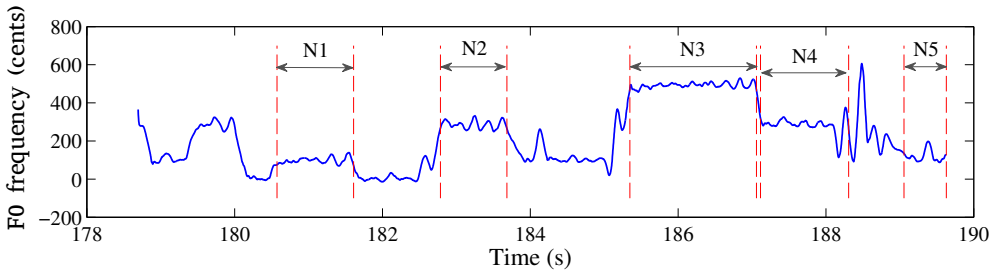

Figure 1. Fragment of a pitch contour showing nyās segments denoted by N_i $(i = 1...5)$

presence of alankārs [2] . The pitch characteristics of a segment depends on the rāg and scale degree of the nyās, and adds further complexity to the task [8]. For example, in Figure 1, N_1 and N_3 have a small pitch deviation from the mean svar frequency, whereas, N_2 and N_4 have significant pitch deviation (close to 100 cents in N_5). Large pitch deviations also pose a challenge in segmentation process. Further, melodic context such as the relative position with respect to a non-voiced or long melodically constant region plays a crucial role in determining a nyās segment. Because of these factors the task of nyās segment detection becomes challenging and requires sophisticated learning techniques together with musically meaningful domain specific features.

In computational analysis of Indian art music, nyās segment detection has not received much attention in the past. To the best of our knowledge, only one study with the final goal of spotting melodic motifs has indirectly dealt with this task [9]. In it, the authors considered performances of a single rāg and focused on a very specific nyās svar, corresponding to a single scale degree: the fifth with respect to the tonic, the 'Pa' svar. This svar is considered as one of the most stable svars, and has minimal pitch deviations. Thus, focusing on it oversimplified the methodology developed in [9] for nyās segment detection.

A related topic is the detection of specific alankārs and characteristic phrases (also referred as *Pakads*) in melodies in Indian art music [10, 11, 12, 13]. These approaches typically exploit pattern recognition techniques and a set of pre-defined melodic templates. A nearest neighbors classifier with a similarity measure based on dynamic time warping (DTW) is a common method to detect patterns in melodic sequences [11, 12]. In addition, it is also the most accurate [14] and extensively used approach for time series classification in general (cf. [15]). Notice that the concept of landmark has been used elsewhere, with related but different notions and purposes. That is the case with time series similarity [16], speech recognition [17, 18], or audio identification [19].

In this paper, we propose a method for detecting occurrences of nyās svar in Hindustani music melodies. The proposed method consists of two main steps: segmentation based on domain knowledge, and segment classification based on a set of musically motivated pitch contour features. There are three main reasons for selecting this approach over a standard pattern detection technique (for example DTW). First, the pitch contour of a nyās segment obeys no explicit patterns, hence the contour characteristics have to be abstracted. Second, information regarding the melodic context of a segment can be easily interpreted in terms of discrete features. Third, we aim to measure the contribution of a specific feature in the overall classification accuracy (for example, if contour variance and length are the most important features for the classification). This is important in order to corroborate the results obtained from such data driven approaches to that from musicological studies.

[2] Characteristic pitch movements acting as ornaments during a svar rendition

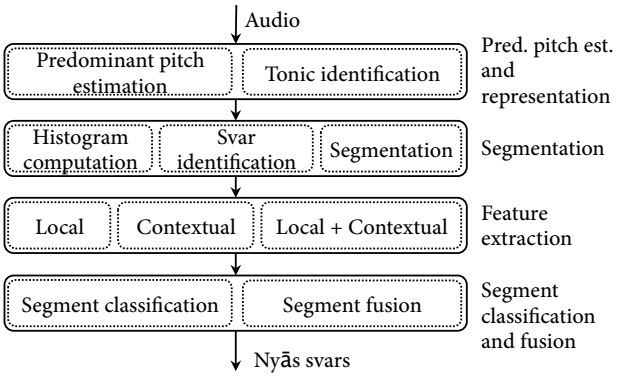

Figure 2. Block diagram of the proposed approach.

The structure of the remainder of the paper is as follows. In Section 2 we present our proposed method for melody segmentation and for detection of nyās segments. In Section 3 we describe our experimental setup, which includes description of the data set and measures used for evaluation, the ground truth annotation procedure, and brief discussion on few baseline methods. In Section 4 we present and discuss the results of our experiments. Finally, in Section 5, we provide some conclusions and directions for future work.

2. METHOD

2.1 Predominant Pitch Estimation & Representation

The proposed method is comprised of four main blocks (Figure 2): predominant pitch estimation and representation, segmentation, feature extraction, and segment classification and fusion. For estimating pitch of the the predominant melodic source [3] we use the method by Sala-

[3] This task is also referred as predominant melody extraction in various contexts within Music Information Research.

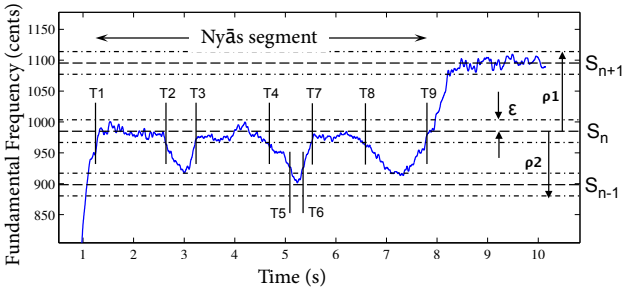

Figure 3. Fragment of a pitch contour containing a nyās segment ($T_1 - T_9$), where T_is denote time stamps and S_ns denote mean svar frequencies. The pitch deviation within the nyās segment ($T_1 - T_9$) is almost 100 cents ($T_5 - T_6$).

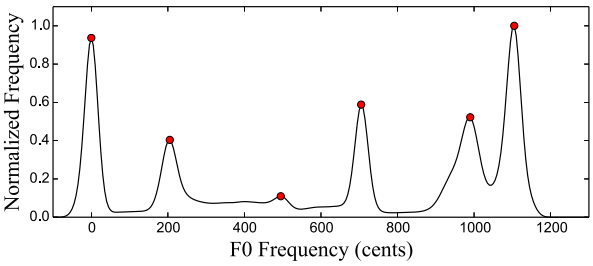

Figure 4. Normalized octave folded pitch histogram used for estimating mean svar frequencies. Estimated mean svar frequencies are indicated by circles.

mon & Gómez [20], which scored very favorably in an international evaluation campaign featuring a variety of musical genres, including Indian art music [4]. For the pitch representation to be musically meaningful, we convert the pitch values to cents, normalized by the tonic frequency of the lead artist. Tonic of the lead artist is extracted automatically using the approach proposed by Gulati [21]. In a comparative evaluation of different tonic identification approaches for Indian art music, this approach consistently performed better for a variety of music material within Indian art music [22]. For both predominant pitch estimation and tonic identification we use the implementations available in Essentia [23], an open-source C++ library for audio analysis and content-based music information retrieval.

2.2 Segmentation

Nyās segment is a rendition of a single svar and the aim of the segmentation process is to detect the svar boundaries. However, svars contain different alankārs as discussed before where pitch deviation with respect to the mean svar frequency can go roughly up to 200 cents. This characteristic of a svar in Hindustani music poses a challenge to segmentation. To illustrate this, in Figure 3 we present an example of a nyās segment (between $T_1 - T_9$, centered around mean svar frequency $S_n = 990$ cents). The pitch deviation in this nyās segment with respect to the mean svar frequency reaches almost 100 cents (between $T_5 - T_6$). Note that the reference frequency, i.e. 0 cent is the tonic pitch of the lead singer.

We experiment with two different methods for segmenting melodies: piece-wise linear segmentation (PLS), a classical, generic approach used for the segmentation of time series data [24], and our proposed method, which incorporates domain knowledge to facilitate the detection of nyās boundaries. For PLS we use a bottom-up segmentation strategy as described in [24]. Bottom-up segmentation methods involve computation of residual error incrementally for each sample of time series. When the residual error satisfies a pre-defined criterion a new segment is created. Out of the two typical criteria used for segmentation, namely average and maximum error, we choose the latter because, ideally, a new segment should be created as soon as the melody progresses from one svar to the other. In order to select the optimal value of the allowed maximum error, which we denote by ϵ, we iterated over four different values and chose the one which resulted in the best performance. Specifically, for $\epsilon = \{10, 25, 50, 75\}$, $\epsilon = 75$ cents yielded the best performance (we rejected $\epsilon \geq 100$ cents in early experimentation stages because few svars of a rāg are separated by an interval of 100 cents and, therefore, the segmentation output was clearly unsatisfactory).

To make the segmentation process robust to pitch deviations, we propose a method based on empirically-derived thresholds. Unlike PLS, our proposed method computes a pitch histogram and uses that to estimate mean svar frequencies before the computation of residual error. This allows us to compute the residual error with respect to the mean svar frequency instead of computing it with respect to the previous segment boundary, as done in PLS. In this way our proposed method utilizes the fact that the time series being segmented is a pitch contour where the values of the time series hover around mean svar frequencies. The mean svar frequencies for an excerpt are estimated as the peaks of the histogram computed from the estimated pitch values. An octave folded pitch histogram is computed using a 10 cent resolution and subsequently smoothened using a Gaussian window with a variance of 15 cents. Only the peaks of the normalized pitch histogram which have at least one peak-to-valley ratio greater than 0.01 are considered as svar locations. As peaks and valleys we simply take all local maximas and minimas over the whole histogram. In Figure 4 we show an example of an octave folded normalized pitch histogram used for estimating mean svar frequencies. The estimated mean svar frequencies are indicated by circles. We notice that the pitch values corresponding to a svar span a frequency region and not a single value.

After we estimate mean frequencies of all the svars in a piece, we proceed with their refinement. For the n-th svar S_n, we search for contiguous segments within a deviation of ε from S_n, that is, $|S_n - P_i| < \varepsilon$, for $i \in [1, N]$, where P_i is the fundamental frequency value (in cents) of the i-th sample of a segment of length N. In Figure 3, this corresponds to segments $[T_1, T_2]$, $[T_3, T_4]$, and $[T_7, T_8]$.

Next, we concatenate two segments $[T_a, T_b]$ and $[T_e, T_f]$ if two conditions are met:

1. $P_i - S_n < \rho_1$ and $S_n - P_i < \rho_2$, for $i \in [T_b, T_e]$,

[4] http://nema.lis.illinois.edu/nema_out/mirex2011/results/ame/indian08/summary.html

where $\rho_1 = S_{n+1} - S_n + \varepsilon$ and $\rho_2 = S_n - S_{n-1} + \varepsilon$.

2. $T_c - T_d < \delta$, where δ is a temporal threshold and $[T_c, T_d]$ is a segment between T_b, T_e such that $|S_m - P_i| < \varepsilon$ for $i \in [T_c, T_d]$ for $m \in [S_{n-1}, S_{n+1}]$ and $m \neq n$.

In simple terms, we concatenate two segments if the fundamental frequency values between them do not deviate a lot (less than ρ_1 and ρ_2) and the time duration of the melody in close vicinity (less than ϵ) of neighboring svars is not too large (less than δ). We repeat this process for all svar locations. In our experiments, we use $\varepsilon = 25$ cents and $\delta = 50$ ms, which were empirically obtained. In the example of Figure 3, we see that the two conditions apply for segments $[T_1, T_2]$ and $[T_3, T_4]$, and not for $[T_3, T_4]$ and $[T_7, T_8]$ because $T_6 - T5 > \delta$. Notice that we can already derive a simple binary flatness measure ν for $[T_a, T_b]$, $\nu = 1$ if $|S_n - P_i| < \epsilon$ for $i \in [T_a, T_b]$ for any n and $\nu = 0$ otherwise.

2.3 Feature Extraction

We extract musically motivated melodic features for segment classification, which resulted out of discussions with musicians. For every segment, three sets of melodic features are computed: local features (L), which capture the pitch contour characteristics of the segment, contextual features (C), which capture the melodic context of the segment, and a third set combining both of them (L+C) in order to analyze if they complement each other. Initially, we considered 9 local features and 24 contextual features:

Local Features: segment length, mean and variance of the pitch values in a segment, mean and variance of the differences in adjacent peak locations of the pitch sequence in a segment, mean and variance of the peak amplitudes of the pitch sequence in a segment, temporal centroid of the pitch sequence in a segment normalized by its length, and the above-mentioned flatness measure ν (we use the average segmentation error for the case of PLS).

Contextual Features: segment length normalized by the length of the longest segment within the same breath phrase [5], segment length normalized by the length of the breath phrase, length normalized with the length of the previous segment, length normalized by the length of the following segment, duration between the ending of the segment and succeeding silence, duration between the starting of the segment and preceding silence, and all the local features of the adjacent segments.

However, after preliminary analysis, we reduced these features to 3 local features and 15 contextual features. As local features we selected length, variance, and flatness measure (ν). As contextual features we selected all of them except the local features of the posterior segment. This feature selection was done manually, performing different preliminary experiments with a subset of the data, using different combinations of features and selecting the ones that yielded the best accuracies.

[5] Melody segment between consecutive breaths of a singer. We consider every unvoiced segment (i.e., a value of 0 in the pitch sequence) greater than 100 ms as breath pause.

2.4 Classification and Segment Fusion

Each segment obtained in Section 2.2 is classified into nyās or non-nyās based on the extracted features of Section 2.3. To demonstrate that the predictive power of the considered features is generic and independent of a particular classification scheme, we employ five different algorithms exploiting diverse classification strategies [25]: trees (Tree), K nearest neighbors (KNN), naive Bayes (NB), logistic regression (LR), and support vector machines with a radial basis function kernel (SVM). We use the implementations available in scikit-learn [26], version 0.14.1. We use the default set of parameters with few exceptions in order to avoid over-fitting and to compensate for the uneven number of instances per class. Specifically, we set `min_samples_split=10` for Tree, `fit_prior=False` for NB, `n_neighbors=5` for KNN, and for LR and SVM `class_weight='auto'`.

For out-of-sample testing we implement a cross-fold validation procedure. We split the data set into folds that contain an equal number of nyās segments, the minimum number of nyās segments in a musical excerpt (7 in our case). Furthermore, we make sure that no instance from the same artist and rāg is used for training and testing in the same fold.

After classification, boundaries of nyās and non-nyās segments are obtained by merging all the consecutive segments with the same segment label. During this step, the segments corresponding to the silence regions in the melody, which were removed during classification, are regarded as non-nyās segments.

3. EXPERIMENTAL SETUP

3.1 Music Collection and Annotations

The music collection used for evaluation consists of 20 audio music recordings of total duration of 1.5 hours, all of which are vocal ālāp performances of Hindustani music. Ālāps are unmetered melodic improvisational sections, usually performed as the opening of a raga rendition. We selected only ālāp performances because the concept of nyās is emphasized in these sections during a rāg rendition. Of the 20 recordings, 15 are polyphonic commercially-available audio recordings compiled as a part of the CompMusic project [6] [27]. The other 5 audio recordings in the data set are monophonic in-house studio recordings of the ālāps sung by a professional singer of Hindustani music. The in-house audio recordings are available under creative commons (CC) license in Freesound [7]. In total we have performances by 8 artists in 16 different rāgs. In order avoid over-fitting of the learned model it is important to include different artists and rāgs as the nyās characteristics highly depend on these aspects [4].

[6] http://compmusic.upf.edu/corpora

[7] http://www.freesound.org/people/sankalp/packs/12292/

Nyās segments were annotated by a performing artist of Hindustani music (vocalist) who has received over 15 years of formal musical training. The musician marked all the nyās segment boundaries and labeled them appropriately. After annotation, we obtained 1257 nyās svar segments. The duration of these segments vary from 150 ms to 16.7 s with a mean of 2.46 s and median of 1.47 s.

3.2 Evaluation Measures and Statistical Significance

For the evaluation of nyās boundary annotations we use hit rates as in a typical music structure boundary detection task [28]. While calculating hit rate, segment boundaries are considered as correct if they fall within a certain threshold of a boundary in the ground-truth annotation. Using matched hits, we compute standard precision, recall, and F-score for every fold and average them over the whole data set. The choice of a threshold however depends on the specific application. Due to the lack of scientific studies on the just noticeable differences of nyās svar boundaries, we computed results using an arbitrary selected threshold of 100 ms. Label annotations are evaluated using standard pairwise frame clustering method as described in [29]. Frames with same duration as threshold value for the boundary evaluation (i.e. 100 ms) are considered while computing precision, recall, and F-score. For assessing statistical significance we use the Mann-Whitney U test [30] with $p < 0.05$ and assuming an asymptotic normal distribution of the evaluation measures. To compensate for multiple comparisons we apply the Holm-Bonferroni method [31], a powerful method that also controls the so-called family-wise error rate. Thus, we end up using a much more stringent criteria than $p < 0.05$ for measuring statistical significance.

3.3 Baselines

Apart from reporting the accuracies for the proposed method and its variants, we compare against some baseline approaches. In particular, we consider DTW together with a KNN classifier ($K = 5$). For every segment, we compute its distance from all other segments and assign a label to it based on the labels of its K nearest neighbors, using majority voting. As the proposed method also exploits contextual information, in order to make the comparison more meaningful, we consider the adjacent segments in the distance computation with linearly interpolated values in the region corresponding to the segment. For comparing with the variant of the proposed method that uses a combination of the local and contextual features, we consider adjacent segments together with the actual segment in the distance computation. As this approach does not consider any features, it will help us in estimating the benefits of extracting musically-relevant features from nyās segments.

In addition, to quantify the limitations of the adopted evaluation measures, we compute a few random baselines. The first one (RB1) is calculated by randomly planting boundaries (starting at 0 s) according to the distribution of inter boundary intervals obtained using the ground-truth annotations. For each segment we assign the labels 'nyās' with a a priory probability (same for all excerpts) computed using ground truth annotations of the whole data set. The second one (RB2) is calculated by planting boundaries (starting at 0 s) at even intervals of 100 ms and assigning class labels as in RB1. Finally, the third one (RB3) considers the exact ground-truth boundaries and assigns the class labels randomly as in RB1 and RB2. Thus, with RB3 we can directly assess the impact of the considered classification algorithms. We found that RB2 achieves the best accuracy and therefore, for all the following comparisons we only consider RB2.

4. RESULTS AND DISCUSSION

We evaluate two tasks, nyās segment boundary annotation, and nyās and non-nyās segment label annotation. For both the tasks, we report results obtained using two different segmentation methods (PLS and the proposed segmentation method), five classifiers (Tree, KNN, NB, LR, SVM), and three set of features (local (L), contextual(C) and local together with contextual (L+C)). In addition, we report results obtained using a baseline method (DTW) and a random baseline (RB2).

In Table 1 we show the results of nyās boundary annotations. First, we see that every variant performs significantly better than the best random baseline. RB2 yields an F-score of 0.184 while the worst variant tested reaches 0.248. Next, we see that the proposed method achieves a notably higher accuracy compared to the DTW baseline. Such difference is found to be statistically significant, with the only exception of the NB classifier. For a given feature set, the performance differences across classifiers are not statistically significant. The only exceptions are Tree and NB, which yield relatively poor and inconsistent performances over different feature sets. We opted to not consider these two classifiers in the following comparisons. Among feature sets, the performance differences are not statistically significant between PLS variants (Table 1, top rows), whereas for the case of the proposed segmentation method (Table 1, bottom rows), we find that the local features perform significantly better than the contextual features and their combination does not yield consistent improvements. Finally, we see that the best results are obtained using the proposed segmentation method together with the local features, with a statistically significant difference to its competitors. Furthermore, the worst accuracy obtained using the proposed segmentation method is notably higher than the best accuracy using PLS method, again with a statistically significant difference.

In Table 2 we show the results for nyās and non-nyās label annotations. Basically, we can draw similar conclusions as with Table 1: (1) all method variants perform significantly better than the random baselines, (2) all the proposed method variants yield significant accuracy increments over the DTW baseline, and (3) no statistically significant differences between classifiers (with the aforementioned exceptions). In label annotations, unlike the boundary annotations, we find that though the local features perform better than the contextual features, the differences are not statistically significant for all the proposed method variants. Furthermore, we also see that the proposed seg-

	Feat.	DTW	Tree	KNN	NB	LR	SVM
A	L	0.356	0.407	0.447	0.248	0.449	0.453
	C	0.284	0.394	0.387	0.383	0.389	0.406
	L+C	0.289	0.414	0.426	0.409	0.432	0.437
B	L	**0.524**	0.672	**0.719**	0.491	**0.736**	**0.749**
	C	0.436	0.629	0.615	**0.641**	0.621	0.673
	L+C	0.446	**0.682**	0.708	0.591	0.725	0.735

Table 1. F-scores for nyās boundary detection using PLS method (A) and the proposed segmentation method (B). Results are shown for different classifiers (Tree, KNN, NB, LR, SVM) and local (L), contextual (C) and local together with contextual (L+C) features. DTW is the baseline method used for comparison. F-score for the random baseline obtained using RB2 is 0.184.

	Feat.	DTW	Tree	KNN	NB	LR	SVM
A	L	**0.553**	0.685	0.723	0.621	0.727	0.722
	C	0.251	0.639	0.631	0.690	0.688	0.674
	L+C	0.389	0.694	0.693	0.708	0.722	0.706
B	L	0.546	**0.708**	**0.754**	0.714	**0.749**	**0.758**
	C	0.281	0.671	0.611	0.697	0.689	0.697
	L+C	0.332	0.672	0.710	**0.730**	0.743	0.731

Table 2. F-scores for nyās and non-nyās label annotations task using PLS method (A) and the proposed segmentation method (B). Results are shown for different classifiers (Tree, KNN, NB, LR, SVM) and local (L), contextual (C) and local together with contextual (L+C) features. DTW is the baseline method used for comparison. The best random baseline F-score is 0.153 obtained using RB2.

mentation method consistently performs better than PLS. However, the differences are not statistically significant.

In addition, we also investigate per-class accuracies for label annotations. We find that the performance for nyās segments is considerably better than non-nyās segments. This could be attributed to the fact that even though the segment classification accuracy is balanced across classes, the differences in segment length of nyās and non-nyās segments (nyās segments being considerably longer than non-nyās segments) can result in more number of matched pairs for nyās segments.

In general, we see that the proposed segmentation method improves the performance over PLS method in both tasks, wherein the differences are statistically significant in the former case. Furthermore, the local feature set, when combined with the proposed segmentation method, yields the best accuracies. We also find that the contextual features do not complement the local features to further improve the performance. However, interestingly, they perform reasonably good considering that they only use contextual information.

5. CONCLUSIONS AND FUTURE WORK

We proposed a method for detecting nyās segments in melodies of Hindustani music. We divided the task into two broad steps: melody segmentation and segment classification. For melody segmentation we proposed a method which incorporates domain knowledge to facilitate nyās boundary annotations. We evaluated three feature sets: local, contextual and the combination of both. We showed that the performance of the proposed method is significantly better compared to a baseline method using standard dynamic time warping based distance and a K nearest neighbor classifier. Furthermore, we showed that the proposed segmentation method outperforms a standard approach based on piece-wise linear segmentation. A feature set that includes only the local features was found to perform best. However, we showed that using just the contextual information we could also achieve a reasonable accuracy. This indicates that nyās segments have a defined melodic context which can be learned automatically. In the future we plan to perform this task on Bandish performances, which is a compositional form in Hindustani music. We also plan to investigate other melodic landmarks and different evaluation measures for label annotations.

Acknowledgments

This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). J.S. acknowledges 2009-SGR-1434 from Generalitat de Catalunya, ICT-2011-8-318770 from the European Commission, JAEDOC069/2010 from CSIC, and European Social Funds.

6. REFERENCES

[1] A. D. Patel, *Music, language, and the brain*. Oxford, UK: Oxford University Press, 2007.

[2] W. S. Rockstro, G. Dyson, W. Drabkin, H. S. Powers, and J. Rushton, "Cadence," in *Grove music online*, L. Macy, Ed. Oxford University Press, 2001.

[3] P. Sambamoorthy, *South Indian music vol. I-VI*. The Indian Music Publishing House, 1998.

[4] A. K. Dey, *Nyāsa in rāga: the pleasant pause in Hindustani music*. Kanishka Publishers, Distributors, 2008.

[5] K. K. Ganguli, "How do we 'see' & 'say' a raga: a perspective canvas," *Samakalika Sangeetham*, vol. 4, no. 2, pp. 112–119, 2013.

[6] 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.

[7] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. A. Murthy, "Classification of Melodic Motifs in Raga Music with Time-series Matching," *Journal of New Music Research*, vol. 43, no. 1, pp. 115–131, Jan. 2014.

[8] S. Bagchee, *Nād understanding raga music*. Business Publications Inc, 1998.

[9] J. C. Ross and P. Rao, "Detection of raga-characteristic phrases from Hindustani classical music audio," in *Proc. of 2nd CompMusic Workshop*, 2012, pp. 133–138.

[10] A. K. Datta, R. Sengupta, N. Dey, and D. Nag, "A methodology for automatic extraction of meend from the performances in Hindustani vocal music," *Journal of ITC Sangeet Research Academy*, vol. 21, pp. 24–31, 2007.

[11] Pratyush, "Analysis and classification of ornaments in North Indian (Hindustani) classical music," Master's thesis, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, 2010.

[12] J. C. Ross, T. P. Vinutha, and P. Rao, "Detecting melodic motifs from audio for Hindustani classical music," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2012, pp. 193–198.

[13] V. Ishwar, S. Dutta, A. Bellur, and H. Murthy, "Motif spotting in an alapana in Carnatic music," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2013, pp. 499–504.

[14] X. Xi, E. J. Keogh, C. R. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," in *Proc. of the Int. Conf. on Machine Learning*, 2006, pp. 1033–1040.

[15] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. J. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275–309, 2013.

[16] C. S. Perng, H. Wang, S. R. Zhang, and D. S. Parker, "Landmarks: a new model for similarity-based pattern querying in time series databases," in *Proc. of the Int. Conf. on Data Engineering (ICDE)*, 2000, pp. 33–42.

[17] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739–1758, 2008.

[18] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative bag-of-visual phrase learning for landmark recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 893–896.

[19] N. Q. K. Duong and F. Thudor, "Movie synchronization by audio landmark matching," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3632–3636.

[20] J. Salamon and E. Gómez, "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, 2012.

[21] S. Gulati, "A tonic identification approach for Indian art music," Master's thesis, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, 2012.

[22] S. Gulati, A. Bellur, J. Salamon, H. Ranjani, V. Ishwar, H. A. Murthy, and X. Serra, "Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation," *Journal of New Music Research*, vol. 43, no. 01, pp. 55–73, 2014.

[23] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an audio analysis library for music information retrieval," in *Proc. of Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 493–498.

[24] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," *Data Mining in Time Series Databases*, vol. 57, pp. 1–22, 2004.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed. Berlin, Germany: Springer, 2009.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] X. Serra, "A multicultural approach to music information research," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2011, pp. 151–156.

[28] B. S. Ong and P. Herrera, "Semantic segmentation of music audio contents," in *Proc. of the Int. Computer Music Conf. (ICMC)*, 2005.

[29] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[30] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[31] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.