TIME-DELAYED MELODY SURFACES FOR RĀGA RECOGNITION

Sankalp Gulati[1] **Joan Serrà**[2] **Kaustuv K Ganguli**[3] **Sertan Şentürk**[1] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Telefonica Research, Barcelona, Spain
[3] Dept. of Electrical Engg., Indian Institute of Technology Bombay, Mumbai, India
sankalp.gulati@upf.edu

ABSTRACT

Rāga is the melodic framework of Indian art music. It is a core concept used in composition, performance, organization, and pedagogy. Automatic rāga recognition is thus a fundamental information retrieval task in Indian art music. In this paper, we propose the time-delayed melody surface (TDMS), a novel feature based on delay coordinates that captures the melodic outline of a rāga. A TDMS describes both the tonal and the temporal characteristics of a melody, using only an estimation of the predominant pitch. Considering a simple k-nearest neighbor classifier, TDMSs outperform the state-of-the-art for rāga recognition by a large margin. We obtain 98% accuracy on a Hindustani music dataset of 300 recordings and 30 rāgas, and 87% accuracy on a Carnatic music dataset of 480 recordings and 40 rāgas. TDMSs are simple to implement, fast to compute, and have a musically meaningful interpretation. Since the concepts and formulation behind the TDMS are generic and widely applicable, we envision its usage in other music traditions beyond Indian art music.

1. INTRODUCTION

Melodies in Hindustani and Carnatic music, two art music traditions of the Indian subcontinent, are constructed within the framework of rāga [3, 29]. The rāga acts as a grammar within the boundaries of which an artist composes a music piece or improvises during a performance. A rāga is characterized by various melodic attributes at different time scales such as a set of svaras (roughly speaking, notes), specific intonation of these svaras, ārōhana-avrōhana (the ascending and descending sequences of svaras), and by a set of characteristic melodic phrases or motifs (also referred to as 'catch phrases'). In addition to these melodic aspects, one of the most important characteristics of a rāga is its calan [23] (literally meaning movement or gait). The calan defines the melodic outline of a rāga, that is, how a melodic transition is made from one svara to another, the precise intonation to be followed during the transition, and the proportion of time spent on each svara. It can also be thought of as an abstraction of the characteristic melodic phrases mentioned above.

Rāga is a core musical concept used in the composition, performance, organization, and pedagogy of Indian art music (IAM). Numerous compositions in Indian folk and film music are also based on rāgas [9]. Despite its significance in IAM, there exists a large volume of audio content whose rāga is incorrectly labeled or, simply, unlabeled. This is partially because the vast majority of the tools and technologies that interact with the recordings' metadata fall short of fulfilling the specific needs of the Indian music tradition [26]. A computational approach to automatic rāga recognition can enable rāga-based music retrieval from large audio collections, semantically-meaningful music discovery, musicologically-informed navigation, as well as several applications around music pedagogy.

Rāga recognition is one of the most researched topics within music information retrieval (MIR) of IAM. As a consequence, there exist a considerable amount of approaches utilizing different characteristic aspects of rāgas. Many of such approaches use features derived from the pitch or pitch-class distribution (PCD) [2, 4, 5, 16]. This way, they capture the overall usage of the tonal material in an audio recording. In general, PCD-based approaches are robust to pitch octave errors, which is one of the most frequent errors in the estimation of predominant melody from polyphonic music signals. Currently, the PCD-based approach represents the state-of-the-art in rāga recognition. One of these approaches proposed by Chordia et al. [2] has shown promising results with an accuracy of 91.5% on a sizable dataset comprising 23 rāgas and close to 550 excerpts of 120 s duration, extracted from 121 audio recordings (note that the authors use monophonic recordings made under laboratory conditions).

One of the major shortcomings of PCD-based approaches is that they completely disregard the temporal aspects of the melody, which are essential to rāga characterization [23]. Temporal aspects are even more relevant in distinguishing phrase-based rāgas [17], as their aesthetics and identity is largely defined by the usage of meandering melodic movements, called gamakas. Several approaches address this shortcoming by modeling the temporal aspects of a melody in a variety of ways [18, 21, 27]. Such approaches typically use melodic progression templates [27], n-gram distributions [18], or hidden Markov models [21]

to capture the sequential information in the melody. With that, they primarily utilize the ārōhana-avrōhana pattern of a rāga. In addition, most of them either transcribe the predominant melody in terms of a discrete svara sequence, or use only a single symbol/state per svara. Thus, they discard the characteristic melodic transitions between svaras, which are a representative and distinguishing aspect of a rāga [23]. Furthermore, they often rely on an accurate transcription of the melody, which is still a challenging and an ill-defined task given the nature of IAM [22, 24].

There are only a few approaches to rāga recognition that consider the continuous melody contour and exploit its raw melodic patterns [6, 11]. Their aim is to create dictionaries of characteristic melodic phrases and to exploit them in the recognition phase, as melodic phrases are prominent cues for the identification of a rāga [23]. Such phrases capture both the svara sequence and the transition characteristics within the elements of the sequence. However, the automatic extraction of characteristic melodic phrases is a challenging task. Some approaches show promising results [11], but they are still far from being perfect. In addition, the melodic phrases used by these approaches are typically very short and, therefore, more global melody characteristics are not fully considered.

In this paper, we propose a novel feature for rāga recognition, the time-delayed melody surface (TDMS). It is inspired by the concept of delay coordinates [28], as routinely employed in nonlinear time series analysis [15]. A TDMS captures several melodic aspects that are useful in characterizing and distinguishing rāgas and, at the same time, alleviates many of the critical shortcomings found in existing methods. The main strengths of a TDMS are:

- It is a compact representation that describes both the tonal and the temporal characteristics of a melody
- It simultaneously captures the melodic characteristics at different time-scales, the overall usage of the pitch-classes in the entire recording, and the short-time temporal relation between individual pitches.
- It is robust to pitch octave errors.
- It does not require the transcription of the melody nor a discrete representation of it.
- It is easy to implement, fast to compute, and has a musically-meaningful interpretation.
- As it will be shown, it obtains unprecedented accuracies in the raga recognition task, outperforming the state-of-the-art by a large margin, without the use of any elaborated classification schema.

In our experiments, we use TDMSs together with a k-nearest neighbor classifier and a set of well known distance measures. The reported results are obtained on two scalable, diverse, and representative data sets of Carnatic and Hindustani music, one of which is originally introduced in this study and made publicly available. To the best of our knowledge, these are the largest publicly available data sets for rāga recognition in terms of the number of recordings, number of rāgas, and total audio duration. The main contributions of the present study are:

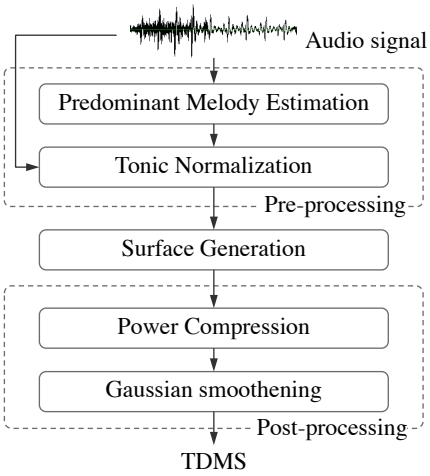

Figure 1. Block diagram for the computation of TDMSs.

- To perform a critical review of the existing methods for rāga recognition and identify some of their main constraints/limitations.
- To propose a novel feature based on delay coordinates, the TDMS, that has all the previously outlined strengths.
- To carry out a comparative evaluation with the best-performing state-of-the-art methods under the same experimental conditions.
- To publicly release a scalable Hindustani music dataset for rāga recognition that contains relevant metadata, annotations, and the computed features.
- To publicly release the code used for the computation of TMDSs and the performed evaluation.

2. RAGA RECOGNITION WITH TIME-DELAYED MELODY SURFACES

2.1 Time-delayed melody surface

The computation of a TDMS has three steps (Figure 1): pre-processing, surface generation, and post-processing. In pre-processing, we obtain a representation of the melody of an audio recording, which is normalized by the tonic or base frequency of the music piece. In surface generation, we compute a two dimensional surface based on the concept of delay coordinates. Finally, in post-processing, we apply power compression and Gaussian smoothing to the computed surface. We subsequently detail these steps.

2.1.1 Predominant melody estimation

We represent the melody of an audio excerpt by the pitch of the predominant melodic source. For predominant pitch estimation, we use the method proposed by Salamon and Gómez [25]. This method performed favorably in MIREX 2011 (an international MIR evaluation campaign) on a variety of music genres, including IAM, and has been used in several other studies for a similar task [7, 12, 13]. We use the implementation of this algorithm as available

in Essentia [1]. Essentia [1] is an open-source C++ library for audio analysis and content-based MIR. We use the default values of the parameters, except for the frame and hop sizes, which are set to 46 and 4.44 ms, respectively. In subsequent steps, we discard frames where a predominant pitch cannot be obtained.

2.1.2 Tonic normalization

The base frequency chosen for a melody in IAM is the tonic pitch of the lead artist [10], to which all other accompanying instruments are tuned. Therefore, for a musically meaningful feature for rāga recognition we normalize the predominant melody of every recording by considering its tonic pitch ω as the reference frequency during the Hertz-to-cent-scale conversion,

$$c_i = 1200 \log_2 \left(\frac{f_i}{\omega} \right),$$

for $0 \leq i < N$, where N is the total number of pitch samples, c_i is the normalized i^{th} sample of the predominant pitch (in cents), and f_i is the i^{th} sample of the predominant pitch (in Hz). The tonic pitch ω for every recording is identified using the multi-pitch approach proposed by Gulati et al. [10]. This approach is reported to obtain state-of-the-art results and has been successfully used elsewhere [8,11]. We use the implementation of this algorithm as available in Essentia with the default set of parameter values. The tonic values for different recordings of an artist are further majority voted to fix the *Pa* (fifth) type error [10].

2.1.3 Surface generation

The next step is to construct a two-dimensional surface based on the concept of delay coordinates (also termed phase space embedding) [15, 28]. In fact, such two-dimensional surface can be seen as a discretized histogram of the elements in a two-dimensional Poicaré map [15]. For a given recording, we generate a surface $\check{\mathbf{S}}$ of size $\eta \times \eta$ recursively, by computing

$$\check{s}_{ij} = \sum_{t=\tau}^{N-1} I\left(B\left(c_t\right), i\right)\ I\left(B\left(c_{t-\tau}\right), j\right)$$

for $0 \leq i, j < \eta$, where I is an indicator function such that $I(x, y) = 1$ iff $x = y$, $I(x, y) = 0$ otherwise, B is an octave-wrapping integer binning operator defined by

$$B(x) = \left\lfloor \left(\frac{\eta x}{1200} \right) \bmod \eta \right\rfloor, \qquad (1)$$

and τ is a time delay index (in frames) that is left as a parameter. Note that, as mentioned, the frames where a predominant pitch could not be obtained are excluded from any calculation. For the size of $\check{\mathbf{S}}$ we use $\eta = 120$. This value corresponds to 10 cents per bin, an optimal pitch resolution reported in [2].

An example of the generated surface $\check{\mathbf{S}}$ for a music piece [2] in rāga Yaman is shown in Figure 2 (a). We see that

[1] https://github.com/MTG/essentia

[2] http://musicbrainz.org/recording/e59642ca-72bc-466b-bf4b-d82bfbc7b4af

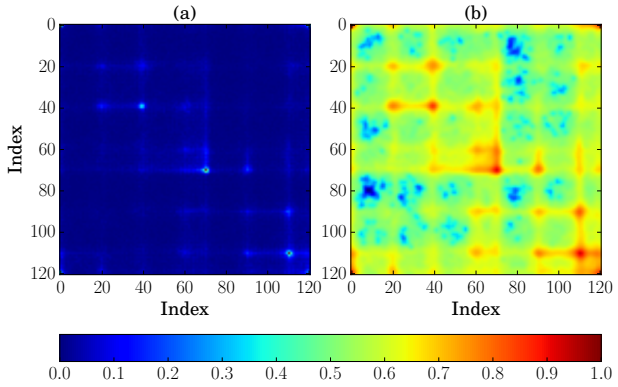

Figure 2. Generated surface for a music piece before (a) and after (b) applying post-processing ($\check{\mathbf{S}}$ and $\hat{\mathbf{S}}$, respectively). For ease of visualization, both matrices are normalized here between 0 and 1.

the prominent peaks in the surface correspond to the svaras of rāga Yaman. We notice that these peaks are steep and that the dynamic range of the surface is high. This can be attributed to the nature of the melodies in these music traditions, particularly in Hindustani music, where the melodies often contain long held svaras. In addition, the dynamic range is high because the pitches in the stable svara regions are within a small range around the svara frequency compared to the pitches in the transitory melodic regions. Because of this, the frequency values in the stable regions are mapped to a smaller set of bins, making the prominent peaks more steep.

2.1.4 Post-processing

In order to accentuate the values corresponding to the transitory regions in the melody and reduce the dynamic range of the surface, we apply an element-wise power compression

$$\overline{\mathbf{S}} = \check{\mathbf{S}}^{\alpha},$$

where α is an exponent that is left as a parameter. Once a more compact (in terms of the dynamic range) surface is obtained, we apply Gaussian smoothing. With that, we attempt to attenuate the subtle differences in $\overline{\mathbf{S}}$ corresponding to the different melodies within the same rāga, while retaining the attributes that characterize that rāga.

We perform Gaussian smoothing by circularly convolving $\overline{\mathbf{S}}$ with a two-dimensional Gaussian kernel. We choose a circular convolution because of the cyclic (or octave-folded) nature of the TDMS (Eqn (1)), which mimics the cyclic nature of pitch classes. The standard deviation of this kernel is σ bins (samples). The length of the kernel is truncated to $8\sigma + 1$ bins in each dimension, after which the values are negligible (below 0.01% of the kernel's maximum amplitude). We experiment with different values of σ, and also with a method variant excluding the Gaussian smoothing (loosely denoted by $\sigma = -1$), so that we can quantify its influence on the accuracy of the system.

Once we have the smoothed surface $\hat{\mathbf{S}}$, there is only one step remaining to obtain the final TDMS. Since the overall duration of the recordings and of the voiced regions within

them is different, the computed surface $\hat{\mathbf{S}}$ needs to be normalized. To do so, we divide $\hat{\mathbf{S}}$ by its L_1 matrix norm:

$$\mathbf{S} = \hat{\mathbf{S}} / \|\hat{\mathbf{S}}\|_1.$$

This also yields values of $\mathbf{S}$, the final TDMS, that are interpretable in terms of discrete probabilities.

The result after post-processing the surface of Figure 2 (a) with power compression and Gaussian smoothing is shown in Figure 2 (b). We see that the values corresponding to the non-diagonal elements are accentuated. A visual inspection of Figure 2 (b) provides several musical insights to the melodic aspects of the recording. For instance, the high salience indices along the diagonal, $(0, 0)$, $(20, 20)$, $(40, 40)$, $(60, 60)$, $(70, 70)$, $(90, 90)$, and $(110, 110)$, correspond to the 7 svaras used in rāga Yaman. Within which, the highest salience at indices (110,110) correspond to the Ni svara, which is the *Vadi* svara, i.e., musically the most salient svara of the rāga, in this case rāga Yaman [23]. The asymmetry in the matrix with respect to the diagonal indicates the asymmetric nature of the ascending and descending svara pattern of the rāga (compare, for example, the salience at indices $(70, 90)$ to indices $(90, 70)$, with the former being more salient than the latter). The similarity of the matrix between indices $(20, 20)$ and $(70, 70)$ with respect to the matrix between indices $(70, 70)$ and $(120, 120)$ delineates the tetra-chord structure of the rāga. Finally, it should be noted that an interesting property of TDMSs is that the mean of the sum across its row and columns yields a PCD representation (see Section 1).

2.2 Classification and distance measurement

In order to demonstrate the ability of the TDMSs in capturing rāga characteristics, we consider the task of classifying audio recordings according to their rāga label. To perform classification, we choose a k-nearest neighbor (kNN) classifier [20]. The reasons for our choice are manifold. Firstly, the kNN classifier is well understood, with well studied relations to other classifiers in terms of both performance and architecture. Secondly, it is fast, with practically no training and with known techniques to speed up testing or retrieval. Thirdly, it has only one parameter, k, which we can just blindly set to a relatively small value or can easily optimize in the training phase. Finally, it is a classifier that is simple to implement and whose results are both interpretable and easily reproducible.

The performance of a kNN classifier highly depends on the distance measure used to retrieve the k neighbors. We consider three different measures to compute the distance between two recordings n and m with TDMS features $\mathbf{S}^{(n)}$ and $\mathbf{S}^{(m)}$, respectively. We first consider the Frobenius norm of the difference between $\mathbf{S}^{(n)}$ and $\mathbf{S}^{(m)}$,

$$D_{\mathrm{F}}^{(n,m)} = \|\mathbf{S}_n - \mathbf{S}_m\|_2.$$

Next, we consider the symmetric Kullback-Leibler divergence

$$D_{\mathrm{KL}}^{(n,m)} = D_{\mathrm{KL}}\left(\mathbf{S}^{(n)}, \mathbf{S}^{(m)}\right) + D_{\mathrm{KL}}\left(\mathbf{S}^{(m)}, \mathbf{S}^{(n)}\right),$$

with

$$D_{\mathrm{KL}}\left(\mathbf{X}, \mathbf{Y}\right) = \sum \mathbf{X} \log\left(\frac{\mathbf{X}}{\mathbf{Y}}\right),$$

where we perform element-wise operations and sum over all the elements of the resultant matrix. Finally, we consider the Bhattacharyya distance, which is reported to outperform other distance measures with a PCD-based feature for the same task in [2],

$$D_{\mathrm{B}}^{(n,m)} = -\log\left(\sum \sqrt{\mathbf{S}^{(n)} \cdot \mathbf{S}^{(m)}}\right).$$

We again perform element-wise operations and sum over all the elements of the resultant matrix. Variants of our proposed method that use D_{F}, D_{KL} and D_{B} are denoted by $\mathcal{M}_{\mathrm{F}}$, $\mathcal{M}_{\mathrm{KL}}$, and $\mathcal{M}_{\mathrm{B}}$, respectively.

3. EVALUATION METHODOLOGY

3.1 Music collection

The music collection used in this study is compiled as a part of the CompMusic project [26]. It comprises two datasets: a Carnatic music data set (CMD) and a Hindustani music data set (HMD). Due to the differences in the melodic characteristics within these two music traditions, and for a better analysis of the results, we evaluate our method separately on each of these data sets. CMD and HMD comprise 124 and 130 hours of commercially available audio recordings, respectively, stored as 160 kbps mp3 stereo audio files. All the editorial metadata for each audio recording is publicly available in Musicbrainz [3], an open-source metadata repository. CMD contains full-length recordings of 480 performances belonging to 40 rāgas with 12 music pieces per rāga. HMD contains full-length recordings of 300 performances belonging to 30 rāgas with 10 music pieces per rāga. The selected music material is diverse in terms of the number of artists, the number of forms, and the number of compositions. In these terms, it can be regarded as a representative subset of real-world collections. The chosen rāgas contain diverse sets of svaras (notes), both in terms of the number of svaras and their pitch-classes (svarasthānās).

Note that CMD has already been introduced and made publicly available in [11]. With the same intentions to facilitate comparative studies and to promote reproducible research, we make HMD publicly available online [4]. Along with the rāga labels for each recording, we also make predominant melody, TDMSs, and the code used for our experiments openly available online.

3.2 Comparison with existing methods

In addition to our proposed method, we evaluate and compare two existing methods under the same experimental setup and evaluation data sets. The two selected methods are the ones proposed by Chordia & Şentürk [2], denoted by $\mathcal{E}_{\mathrm{PCD}}$, and by Gulati et al. [11], denoted by $\mathcal{E}_{\mathrm{VSM}}$. Both approaches have shown encouraging results on scalable

[3] https://musicbrainz.org/

[4] http://compmusic.upf.edu/node/300

datasets and can be regarded as the current, most competitive state-of-the-art in rāga recognition. The former, $\mathcal{E}_{\mathrm{PCD}}$, employs PCD-based features computed from the entire audio recording. The latter, $\mathcal{E}_{\mathrm{VSM}}$, uses automatically discovered melodic phrases and vector space modeling. Readers should note that the experimental setup used in [11] is slightly different from the one in the current study. Therefore, there exists a small difference in the reported accuracies, even when evaluated on the same dataset (CMD). For both $\mathcal{E}_{\mathrm{PCD}}$ and $\mathcal{E}_{\mathrm{VSM}}$, we use the original implementations obtained from the respective authors.

3.3 Validation strategy

To evaluate the performance of the considered methods we use the raw overall accuracy [20]. Since both CMD and HMD are balanced in the number of instances per class, we do not need to correct such raw accuracies to counteract for possible biases towards the majority class. We perform a leave-one-out cross validation [20], in which one recording from the evaluation data set forms the testing set and the remaining ones become the training set. To assess if the difference in the performance between any two methods is statistically significant, we use McNemar's test [19] with $p < 0.01$. To compensate for multiple comparisons, we apply the Holm-Bonferroni method [14]. Besides accuracy, and for a more detailed error analysis, we also compute the confusion matrix over the predicted classes.

In the case of $\mathcal{M}$, a test recording is assigned the majority class of its k-nearest neighbors obtained from the training set and, in case of a tie, one of the majority classes is selected randomly. Because we conjecture that none of the parameters we consider is critical to obtain a good performance, we initially make an educated guess and intuitively set our parameters to a specific combination. We later study the influence of every parameter starting from that combination. We initially use $\tau = 0.3$ s, $\alpha = 0.75$, $\sigma = 2$, and $k = 1$, and later consider $\tau \in \{0.2, 0.3, 0.5, 1, 1.5\}$ s, $\alpha \in \{0.1, 0.25, 0.5, 0.75, 1\}$, $\sigma \in \{-1, 1, 2, 3\}$, and $k \in \{1, 3, 5\}$ (recall that $\sigma = -1$ corresponds to no smoothing; Section 2.1.4).

4. RESULTS AND DISCUSSION

In Table 1, we show the results for all the variants of the proposed method $\mathcal{M}_{\mathrm{F}}$, $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$, and the two state-of-the-art methods $\mathcal{E}_{\mathrm{PCD}}$ and $\mathcal{E}_{\mathrm{VSM}}$, using HMD and CMD data sets. We see that the highest accuracy obtained on HMD is 97.7% by $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$. This accuracy is considerably higher than the 91.7% obtained by $\mathcal{E}_{\mathrm{PCD}}$, and the difference is found to be statistically significant. We also see that $\mathcal{E}_{\mathrm{PCD}}$ performs significantly better than $\mathcal{E}_{\mathrm{VSM}}$. Regarding the proposed variants, we see that, in HMD, $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$ perform better than $\mathcal{M}_{\mathrm{F}}$, with a statistically significant difference.

In Table 1, we see that the trend in the performance for CMD across different methods is similar to that for HMD. The variants $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$ achieve the highest accuracy of 86.7%, followed by $\mathcal{E}_{\mathrm{PCD}}$ with 73.1%. The difference between $\mathcal{M}_{\mathrm{KL}}$ ($\mathcal{M}_{\mathrm{B}}$) and $\mathcal{E}_{\mathrm{PCD}}$ is found to be statistically significant. For CMD, also $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$ perform better than $\mathcal{M}_{\mathrm{F}}$, with a statistically significant difference.

Data set	$\mathcal{M}_{\mathrm{F}}$	$\mathcal{M}_{\mathrm{KL}}$	$\mathcal{M}_{\mathrm{B}}$	$\mathcal{E}_{\mathrm{PCD}}$	$\mathcal{E}_{\mathrm{VSM}}$
HMD	91.3	**97.7**	**97.7**	91.7	83.0
CMD	81.5	**86.7**	**86.7**	73.1	68.1

Table 1. Accuracy (%) of the three proposed variants, $\mathcal{M}_{\mathrm{F}}$, $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{BC}}$, and the two existing state-of-the-art methods $\mathcal{E}_{\mathrm{PCD}}$ and $\mathcal{E}_{\mathrm{VSM}}$ (see text). The random baseline for this task is 3.3% for HMD and 2.5% for CMD.

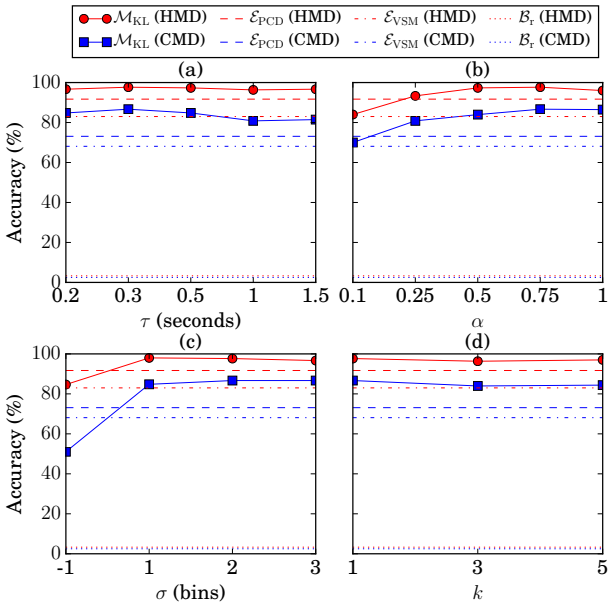

Figure 3. Accuracy of $\mathcal{M}_{\mathrm{KL}}$ as a function of parameter values. State-of-the-art approaches $\mathcal{E}$ and random baselines $\mathcal{B}$ are also reported for comparison.

In general, we notice that, for every method, the accuracy is higher on HMD compared to CMD. This, as expected, can be largely attributed to the difference in the number of classes in HMD (30 rāgas) and CMD (40 rāgas). A higher number of classes makes the task of rāga recognition more challenging for CMD, compared to HMD. In addition to that, another factor that can cause this difference could be the length of the audio recordings, which for HMD are significantly longer than the ones in CMD.

As mentioned earlier, the system parameters corresponding to the results in Table 1 were set intuitively, without any parameter tuning. Since TDMSs are used here for the first time, we want to carefully analyze the influence that each of the parameters has on the final rāga recognition accuracy, and ultimately perform a quantitative assessment of their importance. In Figure 3, we show the accuracy of $\mathcal{M}_{\mathrm{KL}}$ for different values of these parameters. In each case, only one parameter is varied and the rest are set to the initial values mentioned above.

In Figure 3 (a), we observe that the performance of the method is quite invariant to the choice of τ, except for the extreme delay values of 1 and 1.5 s for CMD. This

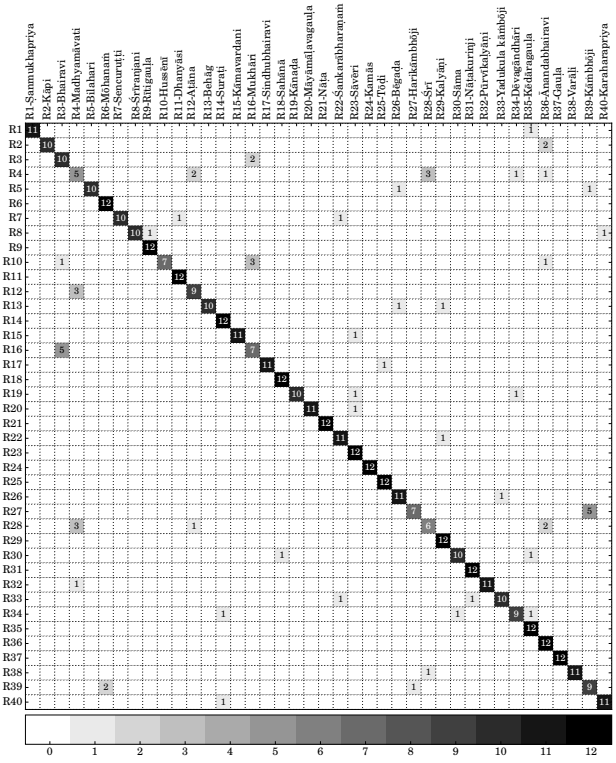

Figure 4. Confusion matrix of the predicted rāga labels obtained by $\mathcal{M}_{\mathrm{KL}}$ on CMD. Shades of grey are mapped to the number of audio recordings.

can be attributed to the melodic characteristics of Carnatic music, which presents a higher degree of oscillatory melody movements and shorter stationary svara regions, as compared to Hindustani music. In Figure 3 (b), we see that compression with $\alpha < 1$ slightly improves the performance of the method for both data sets. However, the performance degrades for $\alpha < 0.75$ for CMD and $\alpha < 0.25$ for HMD. This again appears to be correlated with the long steady nature of the svaras in Hindustani music melodies. Because the dynamic range of $\tilde{\mathbf{S}}$ is high, TDMS features require a lower value for the compression factor α to accentuate the surface values corresponding to the transitory regions in the melodies of Hindustani music. In Figure 3 (c), we observe that Gaussian smoothing significantly improves the performance of the method, and that such performance is invariant across the chosen values of σ. Finally, in Figure 3 (d), we notice that the accuracy decreases with increasing k. This is also expected due to the relatively small number of samples per class in our data sets [20]. Overall, the method appears to be invariant to different parameter values to a large extent, which implies that it is easier to extend and tune it to other data sets.

From the results reported in Figure 3, we see that there exist a number of parameter combinations that could potentially yield a better accuracy than the one reported in Table 1. For instance, using $\tau = 0.3\,\mathrm{s}$, $\alpha = 0.5$, $\sigma = 2$, and $k = 1$, we are able to reach 97.0% for $\mathcal{M}_{\mathrm{F}}$ and 98.0% for both $\mathcal{M}_{\mathrm{KL}}$ and $\mathcal{M}_{\mathrm{B}}$ on HMD. These accuracies are ad-hoc, optimizing the parameters on the testing set. However, and doing things more properly, we could learn the optimal parameters in training, through a standard grid search, cross-validated procedure over the training set [20]. As our primary goal here is not to obtain the best possible results, but to show the usefulness and superiority of TDMSs, we do not perform such an exhaustive parameter tuning and leave it for future research.

To conclude, we proceed to analyze the errors made by the best performing variant $\mathcal{M}_{\mathrm{KL}}$. For CMD, we show the confusion matrix of the predicted rāga labels in Figure 4. In general, we see that the confusions have a musical explanation. The majority of them are between the rāgas in the sets {Bhairavi, Mukhāri}, {Harikāmbhōji, Kāmbhōji}, {Madhyamavatī, Aṭāna, Śrī}, and {Kāpi, Ānandabhairavi}. Rāgas within each of these sets are allied rāgas [29], i.e., they share a common set of svaras and similar phrases. For HMD, there are only 7 incorrectly classified recordings (confusion matrix omitted for space reasons). Rāga Alhaiyā bilāwal and rāga Dēś is confused with rāga Gauḍ Malhār, which is musically explicable as these rāgas share exactly the same set of svaras. Rāga Rāgēśhrī is confused with Bāgēśhrī, which differ in only one svara. In all these cases, the rāgas which are confused also have similar melodic phrases. For two specific cases of confusions, that of rāga Khamāj with Bāgēśhrī, and rāga Darbārī with Bhūp, we find that the error lies in the estimation of the tonic pitch.

5. CONCLUSION

In this paper, we proposed a novel melody representation for rāga recognition, the TDMS, which is inspired by the concept of delay coordinates and Poicaré maps. A TDMS captures both the tonal and the short-time temporal characteristics of a melody. They are derived from the tonic-normalized pitch of the predominant melodic source in the audio. To demonstrate the capabilities of TDMSs in capturing rāga characteristics, we classified audio recordings according to their rāga labels. For this, we used sizable collections of Hindustani and Carnatic music with over 250 hours of duration. Using a k-nearest neighbor classifier, the proposed feature outperformed state-of-the-art systems in rāga recognition. We also studied the influence of different parameters on the accuracy obtained by TDMSs, and found that it is largely invariant to different parameter values. An analysis of the classification errors revealed that the confusions occur between musically similar rāgas that share a common set of svaras and have similar melodic phrases. In the future, we plan to investigate if PCD-based, phrase-based, and TDMSs can be successfully combined to improve rāga recognition. In addition, we would like to investigate the minimum duration of the audio recording needed to successfully recognize its rāga.

6. ACKNOWLEDGMENTS

This work is partly supported by the European Research Council under the European Unions Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

7. REFERENCES

[1] 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. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 493–498, 2013.

[2] P. Chordia and S. Şentürk. Joint recognition of raag and tonic in north Indian music. *Computer Music Journal*, 37(3):82–98, 2013.

[3] A. Danielou. *The ragas of Northern Indian music*. Munshiram Manoharlal Publishers, New Delhi, 2010.

[4] P. Dighe, P. Agrawal, H. Karnick, S. Thota, and B. Raj. Scale independent raga identification using chromagram patterns and swara based features. In *IEEE Int. Conf. on Multimedia and Expo Workshops (ICMEW)*, pages 1–4, 2013.

[5] P. Dighe, H. Karnick, and B. Raj. Swara histogram based structural analysis and identification of Indian classical ragas. In *Proc. of Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 35–40, 2013.

[6] S. Dutta, S. PV Krishnaraj, and H. A. Murthy. Raga verification in Carnatic music using longest common segment set. In *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 605–611, 2015.

[7] S. Dutta and H. A. Murthy. Discovering typical motifs of a raga from one-liners of songs in Carnatic music. In *Int. Soc. for Music Information Retrieval (ISMIR)*, pages 397–402, 2014.

[8] K. K. Ganguli, A. Rastogi, V. Pandit, P. Kantan, and P. Rao. Efficient melodic query based audio search for Hindustani vocal compositions. In *Proc. of Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 591–597, 2015.

[9] T. Ganti. *Bollywood: a guidebook to popular Hindi cinema*. Routledge, 2013.

[10] S. Gulati, A. Bellur, J. Salamon, H. G. Ranjani, V. Ishwar, H. A. Murthy, and X. Serra. Automatic tonic identification in Indian art music: approaches and evaluation. *Journal of New Music Research*, 43(1):55–73, 2014.

[11] S. Gulati, J. Serrà, V. Ishwar, S. Şentürk, and X. Serra. Phrase-based rāga recognition using vector space modeling. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70, 2016.

[12] S. Gulati, J. Serrà, V. Ishwar, and X. Serra. Mining melodic patterns in large audio collections of Indian art music. In *Int. Conf. on Signal Image Technology & Internet Based Systems (SITIS-MIRA)*, pages 264–271, 2014.

[13] S. Gulati, J. Serrà, and X. Serra. An evaluation of methodologies for melodic similarity in audio recordings of Indian art music. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 678–682, 2015.

[14] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70, 1979.

[15] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, Cambridge, UK, 2004.

[16] G. K. Koduri, V. Ishwar, J. Serrà, and X. Serra. Intonation analysis of rāgas in Carnatic music. *Journal of New Music Research*, 43(1):72–93, 2014.

[17] T. M. Krishna and V. Ishwar. Karṇāṭic music: Svara, gamaka, motif and rāga identity. In *Proc. of the 2nd CompMusic Workshop*, pages 12–18, 2012.

[18] V. Kumar, H Pandya, and C. V. Jawahar. Identifying ragas in Indian music. In *22nd Int. Conf. on Pattern Recognition (ICPR)*, pages 767–772, 2014.

[19] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[20] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, USA, 1997.

[21] P. V. Rajkumar, K. P. Saishankar, and M. John. Identification of Carnatic raagas using hidden markov models. In *IEEE 9th Int. Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 107–110, 2011.

[22] S. Rao. Culture specific music information processing: A perspective from Hindustani music. In *2nd CompMusic Workshop*, pages 5–11, 2012.

[23] S. Rao, J. Bor, W. van der Meer, and J. Harvey. *The raga guide: a survey of 74 Hindustani ragas*. Nimbus Records with Rotterdam Conservatory of Music, 1999.

[24] S. Rao and P. Rao. An overview of Hindustani music in the context of computational musicology. *Journal of New Music Research*, 43(1):24–33, 2014.

[25] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

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

[27] S. Shetty and K. K. Achary. Raga mining of Indian music by extracting arohana-avarohana pattern. *Int. Journal of Recent Trends in Engineering*, 1(1):362–366, 2009.

[28] F. Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, 1981.

[29] T. Viswanathan and M. H. Allen. *Music in South India*. Oxford University Press, 2004.