DETECTING MELODIC MOTIFS FROM AUDIO FOR HINDUSTANI CLASSICAL MUSIC

Joe Cheri Ross*, Vinutha T. P.[Ŧ] and Preeti Rao[Ŧ]
Department of Computer Science and Engineering* Department of Electrical Engineering[Ŧ]
Indian Institute of Technology Bombay,
Mumbai 400076, India
joe@cse.iitb.ac.in*
{vinutha,prao}@ee.iitb.ac.in[Ŧ]

ABSTRACT

Melodic motifs form essential building blocks in Indian Classical music. The motifs, or key phrases, provide strong cues to the identity of the underlying *raga* in both Hindustani and Carnatic styles of Indian music. Thus the automatic detection of such recurring basic melodic shapes from audio is of relevance in music information retrieval. The extraction of melodic attributes from polyphonic audio and the variability inherent in the performance, which does not follow a predefined score, make the task particularly challenging. In this work, we consider the segmentation of selected melodic motifs from audio signals by computing similarity measures on time series of automatically detected pitch values. The methods are investigated in the context of detecting the signature phrase of Hindustani vocal music compositions (*bandish*) within and across performances.

1. INTRODUCTION

Hindustani classical music is primarily an oral tradition. While large archives of audio recordings are available, there are few written scores even for widely performed compositions. In such a scenario, retrieval of music based on any relevant high level music descriptors such as *raga* (melodic mode) or *bandish* (a *raga*-specific composition for vocal music) relies entirely on available textual metadata, if any. It is therefore very attractive to consider the automatic extraction of such metadata from audio recordings of concerts. Automatic detection of melodic motifs, for instance, can provide useful inputs to *raga* identification as well as identification of the *bandish* itself [1]. Also, within a concert audio recording, it would be interesting to detect the occurrence of the characteristic melodic phrases thus providing a rich transcription to the listener and the serious student of music. In this work, we consider the problem of detecting specific phrases from recorded performances given one instance of the phrase as template. We also attempt to understand the limitations of the approach in terms of the detection of phrases across performances and artistes.

There is no known previous work on the audio based detection of melodic phrases in Hindustani classical music. A considerable body of recent work, however, has addressed the discovery of melodic patterns from symbolic scores in Western folk music [2]. A continuous-time pitch contour is derived from the score for use in segment alignment and classification. Likewise, the limited reported work on audio signals is based on first obtaining note representations by monophonic pitch transcription [3]. In the case of Hindustani classical music, however, the available symbolic notation is inadequate to deal with tuning variations and complex ornamentation that are fundamentally linked to *raga* characteristics, calling for a different approach to data representation and pattern matching.

In the next section, we review the music background required to appreciate the problem, and outline the challenges. The database and evaluation methods are described next. A framework for the signal processing and pattern matching is proposed. The performance of the system is presented followed by a discussion of the results and prospects for future work.

2. MOTIFS IN HINDUSTANI MUSIC

Hindustani music, especially the modern *khyal* style, is a predominantly improvised music tradition operating within a well-defined *raga* (melodic) and *tala* (rhythmic) framework. Apart from the permitted scale intervals (*swaras*) that define a *raga*, it is its characteristic phrases that complete the grammar and give it a unique identity [4]. Most *ragas* can be adequately represented by up to 8 phrases which then become the essential building blocks of any melody. Thus the detection of recurring phrases can help to identify the *raga*. Indeed musical training involves learning to associate characteristic phrases, or motifs, with ragas. Melodic improvisation involves weaving together a unique melody bound by the chosen rhythmic cycle (*tala*) and consistent with the *raga* phraseology. Often the improvisation is anchored within a known composition, or song, known as *bandish* which provides a platform for exposing a specific *raga*. The *bandish* is identified by its lyrics (especially its title phrase) and melody corresponding to a specific *raga*, *tala* and *laya* (tempo). In the improvised section of the concert known as the *bol-alap*, the singer elaborates within each rhythmic cycle of the *tala* using the words of the *bandish* interspersed with solfege and held vowels, purposefully reaching the strongly accented first beat (the *sam*) of the next rhythmic cycle on a fixed syllable of the signature

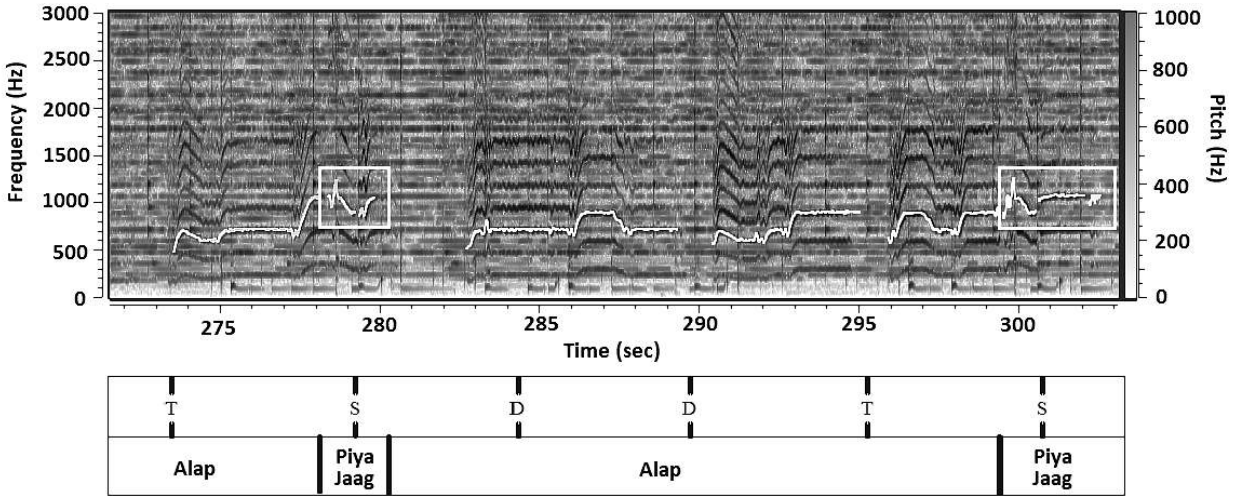

Figure 1. Top: spectrogram with superposed vocal pitch and *mukhda* in boxes; below: first beat of each subcycle (S= *sam*) with aligned lyrics in vocal regions.

phrase of the *bandish*. This recurring phrase, known as the *mukhda,* is the title phrase of the *bandish* and is defined by its text as well as its melodic shape. It acts like a refrain throughout the exposition, which can last several minutes, whereas the other lyrics of the *bandish* can undergo extensive variation in melodic shape in the course of improvisation.

While segmentation of characteristic phrases of the *raga* from a recorded performance is clearly an interesting task that falls within the scope of melodic motif detection, a trained musician is required to notate the recorded performances in order to generate the ground-truth needed for evaluation of any automatic system. On the other hand, the *mukhda* of the *bandish* is easy to segment manually due to the characteristic words of the lyrics and its specific location within the rhythmic cycle. Automatic detection of the *mukhda* can serve to identify the *bandis*h apart from making possible a partial transcription of the performance itself for the interested listener. Although the *mukhda* is detected easily by listening for the lyrics, automatic segmentation cannot rely on such cues due to the known difficulties of speech recognition from singing in polyphonic audio. We focus therefore on the melodic and rhythmic invariances to provide cues for the automatic detection of all occurrences of the *mukhda*, provided one reference instance, across the audio recordings of *bandish* of prominent artistes. Such work can also serve as the basis for more general melodic phrase detection contexts.

3. DATABASE AND EVALUATION METHODS

We selected 4 full-length CD-quality recorded concerts of well-known Hindustani *khyal* vocalists. In all cases, the accompanying instruments are the *tanpura* (drone), *harmonium* and *tabla*. The section of each concert corresponding to *bandish*-based improvisation (*bol-alap*) is extracted for this study. Table 1 shows the artiste names and *bandish* titles with other relevant details including CD cover metadata and the duration of the *bol-alap* section. All the performances use the popular *tintal* rhythm cycle with 16 beats divided equally over 4 sections. The beats are realized by the strokes of the *tabla* (percussion) with the first beat of each section considered to be stressed in 3 of the 4 sections. All the *mukhda* phrases, which may occur around any *sam* (first beat of the cycle) throughout the performance, are manually labeled. This serves as the ground truth ("positives") for the motif detection evaluation. The tempo indicated for each piece is an average, with slow fluctuations in cycle length observed throughout the recordings. Of the four recordings in Table 1, the first two correspond to the same *bandish* by different artistes. The last recording is by a female vocalist. It was observed that this recording with its slow tempo exhibits the largest variations in the duration of the *mukhda* even after accounting for local variations in cycle length.

Artiste	Raga	Tala	Bandish	Tempo (bpm)	Dur. (min)	#Phrases	
						Positive	Negative
Bhimsen Joshi (BJ)	Marwa	Tintal	Guru Bina Gyan	193	4.58	13	55
Ajoy Chakraborti(AC)	Marwa	Tintal	Guru Bina Gyan	205	9.08	33	295
Bhimsen Joshi (BJ)	Puriya	Tintal	Jana na na na	204	9.36	17	97
Kishori Amonkar (KA)	Deshkar	Tintal	Piya Jaag	43	22.3	44	176

Table 1. Description of database

For further processing, the audio is converted to 16 kHz mono at 16 bits/sample. Fig. 1 shows the spectrogram (of the 3 kHz frequency range) of a duration slightly greater than 1 full rhythmic cycle extracted from the *Piya Jaag* recording by Kishori Amonkar. Superimposed on the spectrogram is the detected pitch contour (as obtained by the method presented later in Sec. 4). Beneath the spectrogram is an annotation block depicting the aligned *tala* cycles. The first beat of the cycle is the *sam* (S) corresponds to the *dha* stroke of the tabla. In Fig. 1, the first beat of each sub-cycle is labeled (*dha* (D) or *tha* (T)). The penultimate sub-cycle before the S is the *khali*, as also evident from the absence of low frequency tabla partials in the spectrogram of this segment. The *mukhda* segments corresponding to the utterance "*Piya Jaag*" are enclosed in boxes. The *mukhda* segments are observed to be melodically similar and also similarly aligned within the *tala* cycles. Note that the song syllable that coincides with the *sam* (S) is sometimes left incomplete by the vocalist.

The proposed motif detection method is evaluated in the following experiments. 1) Within-concert detection accuracy where each manually labeled motif serves once as the reference template for all remaining motifs in the same artiste-*bandish* recording; 2) across-concerts detection accuracy where the reference template of a particular artiste is used to find the motifs of the same *bandish* by a different artiste.

4. AUTOMATIC MOTIF DETECTION

The pitch contour depicted in Fig. 1 can be viewed as a time series in which the desired phrase segments are embedded. As such, finding segments in the overall contour that are similar to a given phrase would involve matching the pitch at every time instant of the given phrase to the pitch at every other time instant throughout the time series [3]. It is of interest to explore methods to reduce the search complexity. In the present context, we can exploit the additional knowledge about the rhythmic relationship. As discussed in Sec. 3, the vocalist embeds the *mukhda* phrase in the metrical cycle (*tala*) so that a fixed syllable coincides with the *sam* instant. The metrical space of each cycle is occupied by improvisation culminating with the *mukhda*. Motivated by this, we approach the automatic detection of the *mukhda* phrase from the audio by first identifying a limited set of candidate phrases based on the detected rhythm cycle structure, and then computing a melodic similarity distance between the reference template and each of the candidates.

As in any classification task, it is necessary to design an appropriate data representation and a suitable similarity model for the matching. In this section, we describe the signal processing implementation of a pitch-based data representation and consider similarity models that are suited to the comparison of such time series. Finally, candidate segments with distances from the reference template lower than a threshold are the detected positives.

4.1 Signal Processing

4.1.1 Vocal Pitch Detection

In Hindustani classical vocal music, the accompanying instruments include the drone (*tanpura*), *tabla*, and often, the *harmonium* as well. The singing voice is usually dominant and the melody can be extracted from the detected pitch of the predominant source in the polyphonic mix. Melody detection involves identifying the vocal segments and tracking the pitch of the vocalist. The drone and *harmonium* are strongly pitched instruments. We therefore employ a predominant-F0 extraction algorithm designed for robustness in the presence of pitched accompaniment [4]. This method is based on the detection of spectral harmonics helping to identify multiple pitch candidates in each 10 ms interval of the audio. Next pitch saliency and continuity constraints are applied to estimate the predominant melodic pitch. The best of pitch detection methods achieve no more than 80% accuracy on polyphonic audio. An important factor limiting the accuracy is the fixed choice of analysis parameters, which ideally should be matched to the characteristics of the audio such as the pitch range of the singer and the rate of variation of pitch. In the regions of rapid pitch modulation, characteristic of Indian classical singing, shorter analysis windows serve better to estimate the vocal harmonic frequencies and amplitudes. Hence for better pitch detection accuracy, it is necessary to adapt the window length to the signal characteristics. This is achieved automatically by the maximization of a signal sparsity measure computed at each analysis instance (every 10 ms) for local pitch detection [6]. Finally, it is necessary to identify the vocal regions in the overall tracked pitch. This is achieved by using the peculiar characteristics of Hindustani music where the vocal segments are easily discriminated from the instrumental pitches due to the different temporal dynamics [7].

4.1.2 Motif Candidate Selection

Motivated by the characteristic of the *mukhda*, namely it's alignment with the *sam* stroke of the rhythm cycle, which the artiste pays great importance to achieve, the search algorithm starts by restricting candidate melodic segments to those that match rhythmically. This can achieved via the automatic detection of the beat instants in the audio. In the spectrogram of Fig. 1, the *tabla* strokes corresponding to the beats of the *tala* cycle are visible as vertical impulsive onsets. While the *sam* stroke itself is not particularly distinctive, the *dha* strokes (including the *sam*) can be detected as the highest onsets in the combined energies of two frequency bands: [5000, 8000] and [0, 1000] Hz. The former band is relatively free of interference from vocal partials while the latter band captures the low frequency partial of the *dha* stroke. The filtered output power is subjected to a first-order difference and then half-wave rectified. Spurious peaks are removed by a local threshold. The consistency of the spacing of detected onsets with the known average tempo

is considered further to identify the largest peaks as the *dha* stroke onsets.

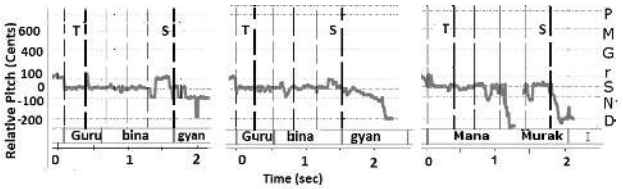

Figure 2. Two positive and one negative phrase of *Guru Bina Gyan*

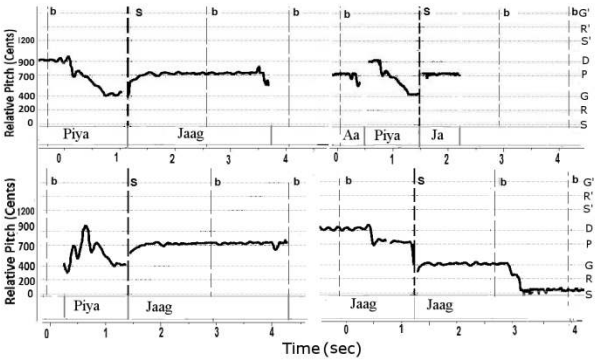

Figure 3. Three positive and one negative (bottom right) phrase of *Piya Jaag*

All audio segments whose alignment around a detected onset matches that of the *mukhda* are treated as potential candidates for motif detection. The extracted segment extends from the instant (*sam*-t1) to (*sam*+t2) where t1 and t2 are nominal values (number of beats in the 16-beat cycle) chosen based on the reference *mukhda* instance. Such a data representation is inherently robust to the slow tempo variations that occur during the concert.

The sequence of pitch values (in cents) obtained across the extracted candidate audio segment is a time series representation that is used further for similarity matching with a reference time series that is similarly obtained. Figures 2 and 3 depict the pitch contours of a few candidate segments showing examples of the melodic and timing variability across *mukhda* realization within concerts. We observe that there are prominent differences in the melodic pitch contour, both in terms of fine pitch variation as well as timing. The note (*swara*) sequence of the *Guru Bina Gyan* phrase is seen to be [Sa, Sa, Ni, Re, Ni, Dha]. However, since the word *Gyan* is often left unsung by the artiste, the *sam* itself serves as the right limit (i.e. t1=5, t2=0) of the *mukhda* in our task. The *swara* corresponding to *Piya Jaag* are [Da, Pa, Ga, Pa]. Here t1=1 and t2=2 were applied to delimit the *mukhda*. Any pitch gaps within the boundaries correspond to pauses. These are filled by linear interpolation or extrapolation of neighbouring pitch values before similarity matching.

4.2 Similarity Modeling

Due to the beat-based method of candidate extraction, the segments tend to be of different absolute durations depending on local tempo variations. Also, singing expressiveness manifests itself in timing changes that can affect the total duration of the sung phrase. The sequence of pitch values obtained at 10 ms intervals throughout the candidate audio segment can be viewed as a non-uniform length time-series representation. We explore two distinct similarity measures for non-uniform length temporal sequences.

Piecewise aggregate approximation has been used to obtain dimension-reduced uniform length time-series for motif discovery in bioinformatics [8]. We apply this method, called SAX, to convert a non-uniform length time series of pitch in cents, computed every 10 ms, to a uniform length, dimension-reduced sequence of symbols. The string length W is varied to determine the optimum dimension of the data representation. A given time-series is aggregated into W uniform length segments each represented by the averaged value of the segment. The real-valued pitch in cents is retained as such but we also consider quantizing pitch to the nearest semitone. Since the tonic frequency is singer dependent in Indian music, the semitone grid is anchored on the most prominent peak of an overall pitch histogram derived from the vocal pitch track across the test audio. Since our present work is confined to within-concert matching, a tonic detection error is inconsequential. Next, the Euclidean distance between the two W-length sequences, the reference and the candidate, is used as a similarity measure.

Another widely used method to compare real-valued time series related to each other through, possibly, nonlinear time-scaling, is the dynamic time-warping (DTW) distance measure [9]. The distance between the so aligned reference and candidate phrases is used as the similarity measure. Pathological warpings are avoided by incorporating the Sakoe-Chiba constraint on the width of a diagonal band in the DTW path matrix. The absolute difference in cents between pitch values is used as the local distance in the DTW path optimization. Any absolute difference within 25 cents (i.e. a quarter tone) is rounded down to 0 cents. This is found to help reduce the influence of practically imperceptible pitch differences on the warping path and therefore any unnecessary stretching of the path.

5. EXPERIMENTS AND DISCUSSION

Given the database described Table 1, we evaluate the different data representations and similarity measures on within-concert and across singer-concert motif detection tasks. Each candidate phrase extracted from the detected onsets as presented in Sec. 4 is labeled positive or negative depending on whether or not it is the actual motif (i.e. *mukhda* phrase). Table 1 shows the number of such phrase segments available for the evaluation of the motif detection methods. To maximize the use of the available annotated data, each labeled motif is considered as the reference once with all other motifs serving as positive tokens and the remaining candidates as negative tokens. Thus, the *Piya Jaag* motif detection task can be evaluated on 44x43 = 1892 positive pairs and 44x176 = 7744 negative pairs (i.e. each positive with all negatives). Table 2 summarizes the experiments. The Experiment A considers motif detection from within the *Guru Bina* recording of Bhimsen Joshi given a reference template from the same recording. Similarly, the Experiments B, C and D

consider the within-concert detection as specified in Table 2. The Experiment E uses the positive tokens of *Guru Bina* by BJ to detect the *mukhda* in the same *bandish* concert by a different vocalist, AC. As it turns out, the two male singers are tuned to the same tonic. In each experiment, the rate of false alarms for a given hit rate (correct detections) is computed for each combination of similarity model and data representation. The similarity measures include SAX and DTW. The data representations chosen for the study are either the continuous pitch values (i.e. 1200 cents per octave) indicated by "q1200", or the quantized versions (12 semitones per octave on an equitempered scale) indicated by "q12".

Fig. 4 shows an example of the distribution of distances for positive-positive pairs and positive-negative pairs. The recording is *Piya Jaag* (Experiment D) evaluated with DTW-q1200. We observe that the distances between the positive phrases cluster closely relative to the distances between the positive-negative phrase pairs. There is a limited overlap between the two distributions. That the spread of the negative distances is relatively wide indicates the robustness of the distance measure in terms of its discrimination of melodic shapes. We also note the presence of a small isolated cluster of positive distances. A closer examination revealed that this stemmed from the wide timing variability across *Piya Jaag* phrases with its particularly slow tempo. Thus there were at least two distinct behaviours within the set of positive phrases. The inter-phrase distances between the longer duration phrases tended to be lower than the distances involving shorter duration phrases. Fig. 5 shows the ROC (hit rate versus false alarm rate) derived from the distributions of Fig. 4 by varying the decision threshold. We observe two bends in the curves, consistent with the bimodal shape of the pdf.

Table 3 summarizes the classification results across the experiments in terms of false alarm rate (FA) for a fixed HR chosen near the knee of the ROCs of the corresponding data. Given that the extracted candidate phrases have durations varying in the range of 2-4 sec (200-400 length string), we vary the SAX string length around W=50 (corresponding to the aggregation of 4-8 samples). Preliminary experiments revealed that W substantially lower than this (i.e. more averaging) led to worsened performance. We note that the performance of SAX improves with pitch quantization at a fixed string length of 50. Increasing or decreasing the string length around this does not improve performance on the whole. The DTW system performs substantially better than SAX in terms of reducing the FA at fixed hit rate. As in the case of SAX, pitch quantization helps to improve performance further in some cases. That DTW does relatively better indicates that non-uniform time warping is essential to achieve the needed alignment between phrases before distance computation. This is consistent with what is known about the tradition where the essential melodic sequence of the *mukhda* phrase is strongly adhered to by the singer while metrical alignment is focused only on getting to the specific syllable onset (e.g. *Gyan* in Fig. 2 and *Jaag* in Fig. 3) on the first beat of the cycle (the *sam*).

Expt	Bandish	Singer	#Phrases	
			POS	NEG
A	Guru Bina	BJ	156	715
B	Guru Bina	AC	1056	9735
C	Jana na na na	BJ	272	1649
D	Piya Jaag	KA	1892	7744
E	Guru Bina	BJ vs AC	429	3835

Table 2. Description of experiments with number of positive and negative phrase candidates available in each

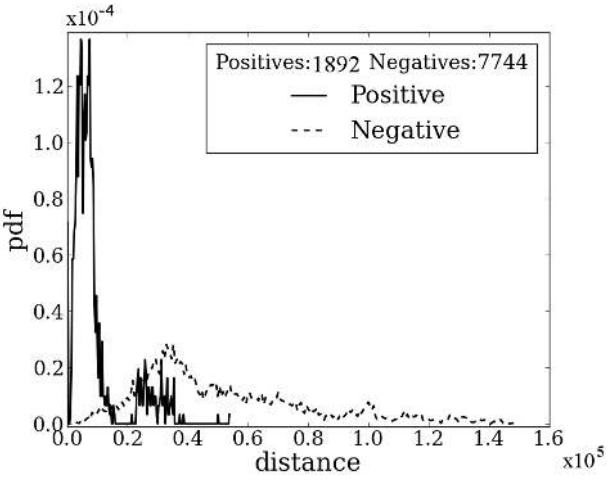

Figure 4. DTW distances distribution for *Piya Jaag* recording

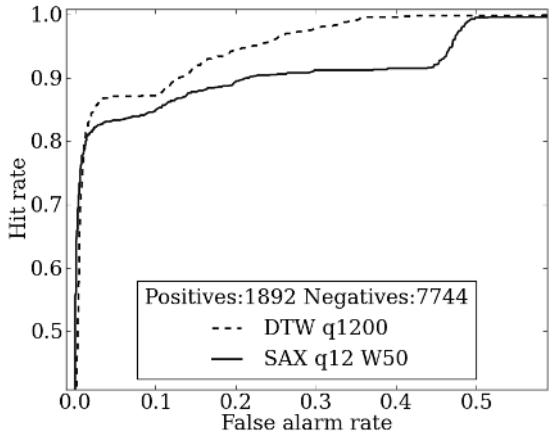

Figure 5. ROC curves for *Piya Jaag* distribution

Further, comparing the Experiments E and B as a case of between-concert to within-concert performance of the motif detection methods, we see that the FA is somewhat higher in Experiment E which involves a reference motif

Method	Experiment A		Experiment B		Experiment C		Experiment D		Experiment E	
	HR	FA	HR	FA	HR	FA	HR	FA	HR	FA
SAX-q1200 -W50	1	.096	.94	.019	.86	.239	.87	.130	.94	.035
SAX-q12-W40	1	.084	.94	.016	.86	.231	.87	.135	.94	.024
SAX-q12-W50	1	.071	.94	.015	.86	.216	.87	.124	.94	.029
SAX-q12-W60	1	.091	.94	.014	.86	.210	.87	.133	.94	.023
DTW-q1200	1	.044	.94	.007	.86	.044	.87	.032	.94	.015
DTW-q12	1	.053	.94	.008	.86	.042	.87	.025	.94	.014

Table 3. Performance of SAX and DTW motif detection under different configurations. WX = SAX string dimension is X; qY= quantized pitch levels per octave; HR = hit rate; FA = number of false alarms

from the concert of the same *bandish* by a different artiste. This is consistent with the anticipated higher variability in motif contour across artistes.

6. CONCLUSION

Similarity measures traditionally used in time-series matching have been shown to perform well in the context of melodic motif detection in the improvised *bandish* of Hindustani vocal concert recordings. The processing of the polyphonic audio signals needed to achieve a suitable data representation was presented. Musical knowledge related to the metrical relation between the *mukhda* motif and the underlying rhythmic structure was exploited to achieve a reduced search space, using available similarity measures, and possibly more robust detection. While the *mukhda* context considered in this work is relevant in both Hindustani and Carnatic vocal music (in the *bol-alap* and *niraval* respectively), the detection of other characteristic *raga* phrases would be a logical extension. It is not clear whether rhythmic cues would help in this more general melodic segmentation. Further extension to unsupervised clustering of phrases in a concert recording can contribute to higher-level classification tasks such as *raga* recognition as well to further research in audio transcription for such musical traditions.

7. ACKNOWLEDGEMENT

This work received partial funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 267583 (CompMusic).

8. REFERENCES

[1] J. Chakravorty, B. Mukherjee and A. K. Datta: "Some Studies in Machine Recognition of Ragas in Indian Classical Music," *Journal of the Acoust. Soc. India*, Vol. 17, No.3&4, 1989.

[2] Z. Juhasz: "Analysis Of Melody Roots In Hungarian Folk Music Using Self-Organizing Maps With Adaptively Weighted Dynamic Time Warping," *Journal Applied Artificial Intelligence*, Vol.21, No.1, 2007.

[3] R. B. Dannenberg and N. Hu: "Pattern Discovery Techniques for Music Audio," *Journal of New Music Research*, Vol. 32, No.2, 2002.

[4] S. Rao, W. van der Meer and J. Harvey: "The Raga Guide: A Survey of 74 Hindustani Ragas," Nimbus Records with the Rotterdam Conservatory of Music, 1999.

[5] V. Rao and P. Rao: "Vocal Melody Extraction in the Presence of Pitched Accompaniment in Polyphonic Music," *IEEE Trans. Audio Speech and Language Processing*, Vol. 18, No.8, 2010.

[6] V. Rao, P. Gaddipati and P. Rao: "Signal-driven Window-length Adaptation for Sinusoid Detection in Polyphonic Music," *IEEE Trans. Audio, Speech, and Language Processing,* Vol. 20, No.1, 2012.

[7] V. Rao, C. Gupta and P. Rao: "Context-aware Features for Singing Voice Detection in Polyphonic Music," *Proc. of Adaptive Multimedia Retrieval*, 2011.

[8] J. Lin, E. Keogh, S. Lonardi and B. Chiu: "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," In *Proc. of the Eighth ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.

[9] D. Berndt and J. Clifford: "Using Dynamic Time Warping to Find Patterns in Time Series," *AAAI-94 Workshop on Knowledge Discovery in Databases,* 1994.