To appear in Frontiers of Research on Speech and Music (FRSM)
November 11-12, 2016, Baripada, India.

IMPROVED MELODIC SEQUENCE MATCHING FOR QUERY BASED SEARCHING IN INDIAN CLASSICAL MUSIC

Ashwin Lele[#], Saurabh Pinjani[#], Kaustuv Kanti Ganguli, and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology Bombay, Mumbai, India.
{ashwinlele,saurabh.pinjani,kaustuvkanti,prao}@ee.iitb.ac.in[1]

Abstract- *Time-series pattern matching techniques have recently been extensively applied to the task of melodic phrase retrieval from audio for Indian classical vocal music. The conversion to a relatively sparse representation by retaining only flat regions of the pitch contour and introducing domain driven cost functions in the string search have been reported to show some reduction in retrieval accuracy while significantly reducing computational complexity. A combination of knowledge- and data-driven optimization on a database of commercial raga performance audios is proposed to counter the accuracy loss. Two different melodic representations are tested on mukhda phrase detection in a search-by-query framework. We present results that demonstrate improvement in retrieval accuracy on incorporating duration information of the transcribed notes. Further, query dependent preset parameter settings for the subsequence search are shown to help.*

1. INTRODUCTION

Time sequence alignment has consistently been a problem of active interest in the field of music information retrieval and the dynamic time warping (DTW) algorithm has been successfully implemented for Hindustani classical music A DTW based distance measure was successfully used on the time-series segments to model melodic similarity on a raga-specific corpus [1]. This, though efficient, is computationally heavy which proves to be a barrier in the way of its application to large dataset. In order to make the task computationally efficient a transcription that converts the audio to a string of only steady notes was implemented and the results are found to be comparable with that of DTW baseline system [2]. With the motivation of identifying musical repetitions [3], the mukhda phrases of a particular composition (bandish) that are characteristic of a raga have been used as queries [4] in a phrase retrieval framework. This task if tuned to optimum efficiency, will prove to be very efficient in the following applications:

- Searching for a particular song on the basis of only characteristic phrases from a large dataset.
- This method can be extended to other music genres where the stable note regions lend themselves to quantization.

String-matching techniques such as the Smith-Waterman algorithm have previously been applied to find the approximate longest common subsequence between the query and target segments of discrete symbols [2]. In this paper we analyse the effects of different transcriptions on the performance of the subsequence search algorithm. We include the durations of the transcribed notes to the cost function as an additional knowledge-driven cue. The areas in which this work contributes are:

- We provide insight on the effect of transcription on retrieval accuracy.
- We propose query dependent preset parameter settings for the subsequence search algorithm that predicts 'difficult' queries characterized by heavily ornamented phrases.

The structure of the paper is as follows. The following section discusses relevant musical concepts, scoring scheme of the standard Smith-Waterman algorithm and evaluation metrics. Section 3 analyses the effect of time information inclusion on the performance with respect to the previously reported results [2]. In the final section we summarise our findings and propose planned research directions that intuitively follow.

[1] The authors marked with # contributed equally to this work.

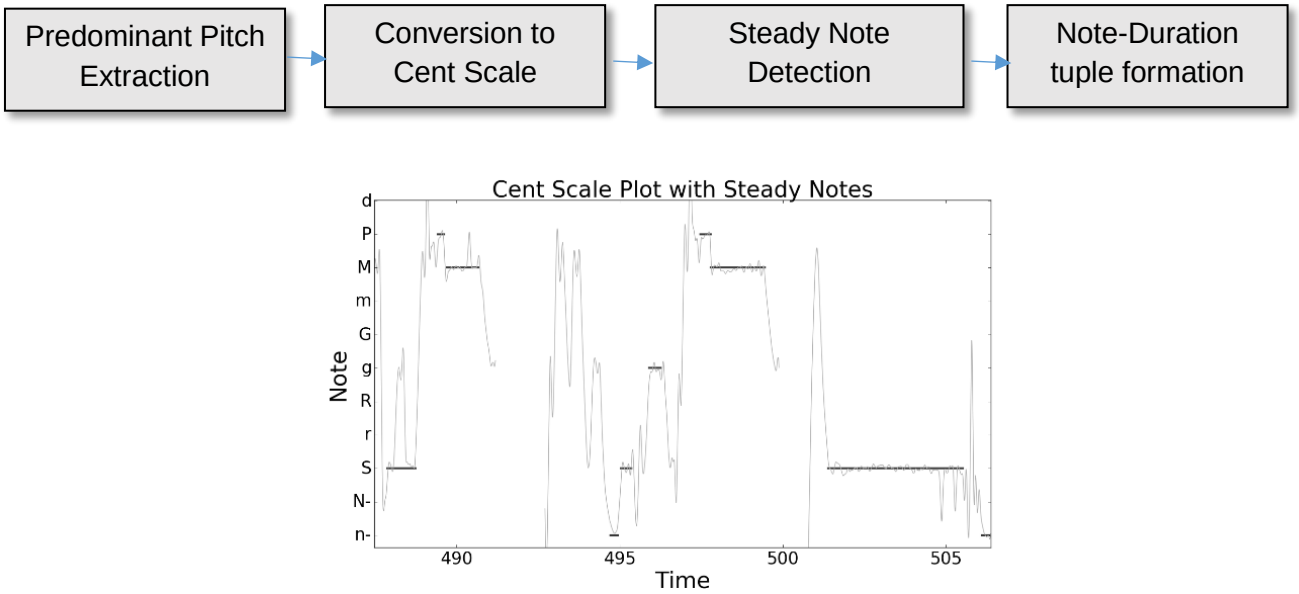


Figure 1. Pitch contour of an audio snippet with the detected steady notes marked by dark horizontal lines. The block diagram of the overall processing is shown on top.

2. BACKGROUND

Hindustani classical music composition (bandish) in a raga is characterized by a mukhda which is a cyclically occurring phrase with a specific location in the tala cycle and having particular lyrics. This phrase is recognized by a trained musician on the basis of note sequences, position in tala cycle and its lyrics with decreasing level of difficulty. Automatic detection of mukhda phrases in a raga can be thought of as an extended raga detection problem as these mukhda phrases are closely associated with the raga and serve as the most basic tools in raga recognition. The following procedure was implemented to convert the audio to a different representation for computation [2].

Transcription Baseline System: The transcription flowchart for this algorithm is shown in Figure 1. First the audio is converted to a string of fundamental frequency of vocal audio using F0 detection. For this a vocal audio pitch extraction from the polyphonic audio containing tabla, tanpura and harmonium or sarangi predominant fundamental frequency detection is implemented by the salience based combination of two algorithms [5] which exploit the spectral properties of the voice with temporal smoothness constraints on the pitch. Then, the segments in the audio where no vocal audio is detected, zero value to assign to the pitch. The other frequencies are converted to the cent scale and the steady notes in the audio are detected using an algorithm [2]. Finally a tuple of steady note and its duration is created which is given as output by the transcription algorithm.

This being the basic procedure there are two varieties of transcription implemented to judge the effect the improved algorithm will have. These two are explained below and will be used to see the effect a transcription system has on the performance of SW algorithm.

Transcription 1: If a silence or a transition is preceded and succeeded by the same steady note these two notes are joined. This is treated as a single note with the duration as the sum of duration of these notes in the audio. The strategies are described in the chart shown below.

Transcription 2: All the steady notes that are detected are assigned the durations they have in the audio.

Transcription 1						
Note	Sa	Ni	Ga	Sa	Ma	Sa
Duration	6.40438	0.377774	0.284441	1.364431	0.297774	1.51554

Transcription 2								
Note	Sa	Ni	Ga	Sa	Sa	Ma	Sa	Sa
Duration	6.40438	0.377774	0.284441	0.408885	0.955546	0.297774	0.55555	0.95999

2.1. Smith Waterman Algorithm

The Smith Waterman algorithm[6] is based on a scoring system which allots two kinds of scores to every possible match between the query and candidate. Finally the candidates having a score above a particular value are selected. These two scores are namely Similarity penalty and Gap penalty.

2.1.1. The Scoring and Evaluation System

Similarity score: The comparison between the notes of query and candidate is performed in one-to-one fashion. A high positive score for an exact match between the note in query and in the candidate phrase. A slightly smaller score is assigned when the notes are less than 2 semitones apart. A negative score is assigned when the notes are farther than this threshold.

Gap Function: The candidate phrases may contain a string of notes within the query that is completely dissimilar compared to the query and this subsequence is termed as gap. The penalty system was chosen to penalize proportional to the number of notes in the gap. The detailed description can be found in [2].

$$Precision = \frac{number\ of\ true\ hits}{\text{total number of detections}} \qquad Recall = \frac{number\ of\ true\ hits}{\text{total number of ground truths}}$$

A detection is termed as a true hit if there is at least 50% overlap between the candidate and the query in the time scale. The ground truths are generally from the characteristic phrases of the raga called pakad and impose a heavy restriction on exact searching as these phrases occur with small variation throughout the recital. The second major challenge in this detection procedure is the use of 'tihayi' which increases as the recital progresses where a single phrase is repeated thrice. In most cases it has the note sequence same as the mukhda phrase and therefore these detection are false alarms because they match with the query though are not mukhda phrases actually. We attempt to overcome these issues with the improvements in the algorithm that we explain subsequently in this paper.

2.2. Dataset

The dataset consists of 75 commercially available recordings of 25 artists in case of transcription-2. For transcription-1 subset of the above dataset of 41 songs by 18 artists is utilized. The recordings are carefully selected to take into account different singing styles and ragas of different nature. Please refer [2] for detailed description of the dataset utilised.

3. TIME INFORMATION INCLUSION

The existing algorithm for motif detection was a direct adaptation of the Smith Waterman Algorithm used for DNA sequencing where the nucleotides are analogous to the musical notes. However in the case of music, each note also has with it its corresponding duration. Hence Smith Waterman algorithm must be suitably modified to handle note duration.

But the primary question is that how important is the note duration information. Does the duration of the notes in a motif affect its musical meaning? It is observed that note duration is critical to the perception of musical phrases. This means that if we want the motif detection system to provide us more musically meaningful matches, duration of notes has to be taken into account. Therefore it is necessary to make some modifications to the Smith Waterman Algorithm.

To achieve this task we define a quantity called "fraction". It is defined as the ratio of the duration of the two notes being compared.

$$Fraction = max\left\{\frac{T_{Query}}{T_{Song}}, \frac{T_{Song}}{T_{Query}}\right\}$$

Where T_{Query} and T_{Song} are the durations of the note in the query and song respectively

This quantity by the nature of its definition is greater than or equal to 1.This property proves to be useful in designing the penalty scheme. If fraction is very large means that the notes have large difference in their duration and hence they must be musically very different. Therefore a penalty must be imposed. However we must also note that minor variations may occur in the duration of notes due to the uniqueness in the style of every singer. The value of fraction will deviate slightly from 1. However it is best that such minor variations be ignored so as to accommodate the uniqueness and originality of each artist. So there should be a threshold value of fraction under which minor aberration in fraction are to be ignored. We have empirically taken this threshold value to be 1.25.

3.1. Modified Penalty Scheme

Similarity Function with Time Inclusion

The fact that the value of fraction is greater than or equal to one is used in the penalty scheme.

- *Same Score*: It is positive. For a large value of fraction one must reduce this score. Hence it is divided by fraction.
- *Far Penalty:* It is negative and hence its absolute value must increase with increase in the value of fraction. Hence it is multiplied with fraction.
- *Gap Margin*: As the value of fraction increases our tolerance for difference in must reduce. Hence gap is divided by fraction.

Gap Function with Time Inclusion

The value of the penalty for gaps is also modified to incorporate time. Previously the algorithm based on the penalty only on the number of missing/additional notes that constitute the gap. However due to high ornamentation the transcription leads to occurrence of many notes with small duration. To handle this we multiply the number of notes in the gap into the sum of their duration. In such a case the duration of the note acts as a failsafe by removing the emphasis of the extraneous notes that have crept in due to high ornamentation. Also intuitively a gap with longer duration makes the query dissimilar to the original motif. Hence penalty must be imposed due to this as well.

$$Gap\ Penalty = -[\, m * (Number\ of\ notes\ in\ gap\ ) * (duration\ of\ gap) + c\ ]$$

The gap penalty parameters used earlier were 0.8, 0.8. Various values of these parameters were simulated and the performance was noted. The parameters were finally tuned such that maximum precision and recall pair was obtained.

4. RESULTS

The formalism developed was applied to both the transcriptions mentioned above. For comparing between the schemes including time and without inclusion of time we define new term called good query. The annotated ground truths are used as queries. For any query if the recall is above 0.3 we define it as a 'good query'. The fraction of good queries in a song is then defined as the ratio of number of good queries to the total number of queries. The results are summarized in the Figure 3**.** The fraction of good queries increase with the incorporation of time in most songs. The 43 songs shown in the diagram are those for which there exists at least 1 good query

Figure 2 depicts the effect of time information inclusion on performance metrics precision and recall. For transcription -II there is consistent improvement in recall. Precision also improves for 80% of songs. In the case of the transcription-I the improvement is inconsistent. This is because we are combining the notes if they are same and the originality of the song is lost. The incorporation of duration is now not serving as a value addition

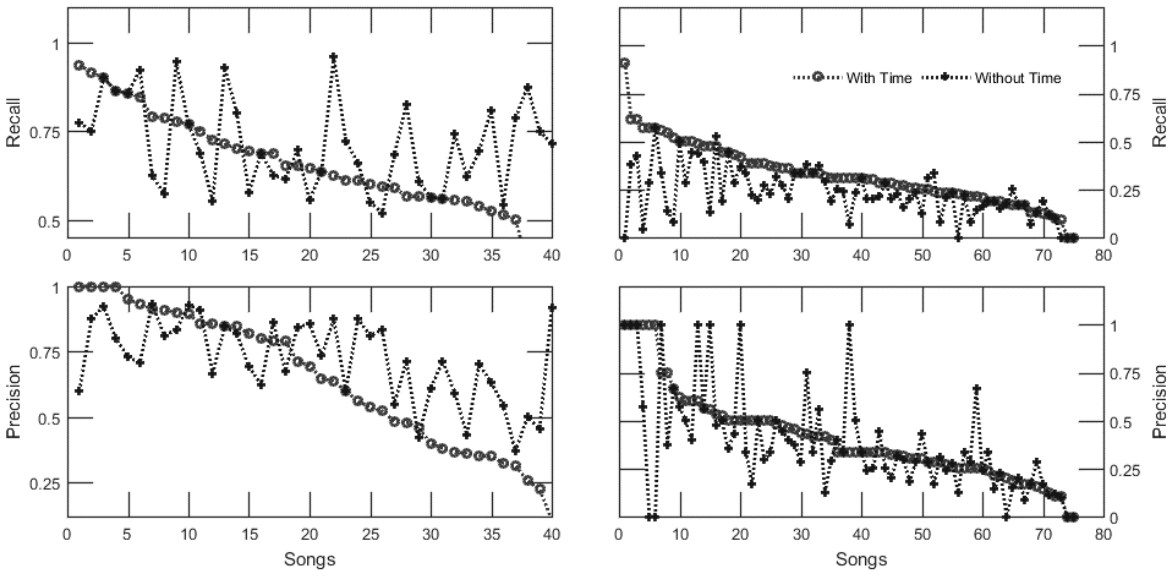


Figure 2. Effect of time inclusion for both transcription schemes. Note that the axes ranges have been kept different to provide clear insight.(The left and right columns represents transcription-1 and transcription-2 respectively).

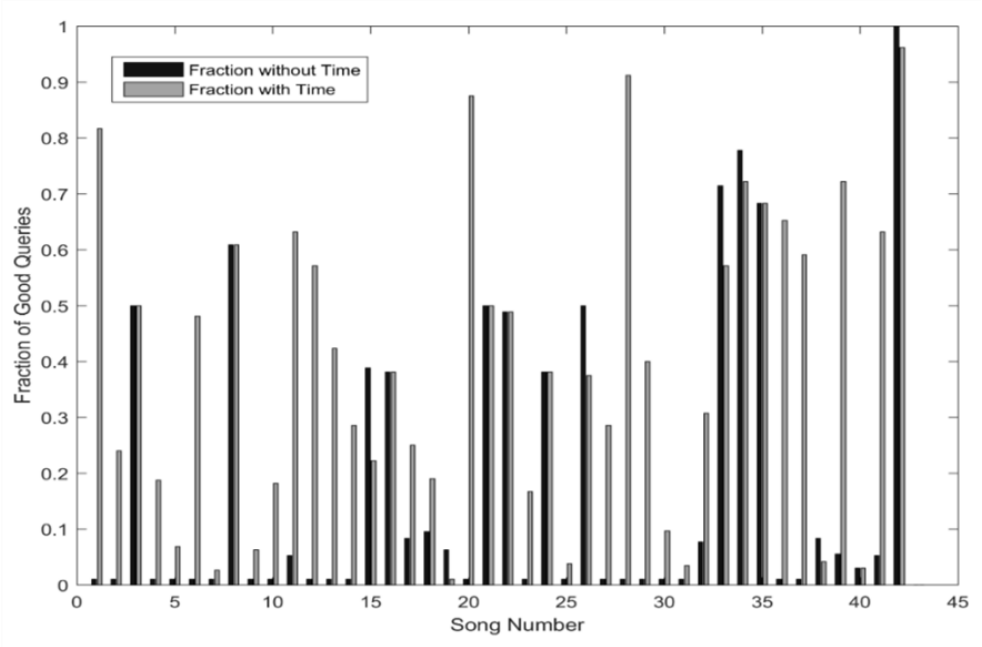


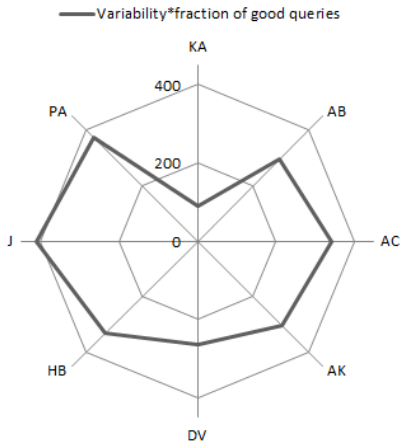
Figure 3. Improvement of performance due to incorporation of duration information.

4.1. Predictive Modelling with Variability

Variability [2] is defined a

$$Variability = \frac{(Number\ of\ Notes\ in\ Query)^2}{Total\ Duration\ of\ Query}$$

The variability therefore is proportional to the amount of ornamentation around the notes that a singing style has. This analysis was carried out for eight artists for all the ground truths in their recordings and average variability for each artist's queries was calculated. The results are summarised in Figure 4.



KA – Kishori Amonkar
AB – Ashwini Bhide Deshpande
AC – Ajoy Chakraborty
AK – Amir Khan
DV – Dattatray Velankar
HB – Hirabai Barodekar
J – Jasraj
PA – Prabha Atre

Figure 4. Variabilities and retrieval efficiency for eight singers is being compared. The product of fraction of good queries and variability (denoted by the radius of the graph) is seen to remain almost constant. This means that the retrieval efficiency in lower for highly ornamented phrases.

5. CONCLUSION

We developed a modified Smith-Waterman algorithm to include the durations of the notes. The method was tested on two different transcription schemes and the results were summarized showing improved retrieval performance. A study to find correlation between singing style and retrieval performance was carried out. Following are the topics in which future work may be carried out to optimize the string searches even further:

- The effect of transcription methodology on the retrieval efficiency has to be studied rigorously.
- We plan to develop a framework to improve retrieval for recordings with high variability.
- Until now we have considered only the steady note regions in the melodic transcription and the transitions between the notes have been completely ignored. A modified framework that also models transient segments is worth investigating further.

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).

References

[1] 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," *J. New Music Res.*, vol. 43, no. November 2014, pp. 115–131, 2014.

[2] K. K. Ganguli, A. Rastogi, V. Pandit, P. Kantan, and P. Rao, "Efficient Melodic Query Based Audio Search for Hindustani Vocal Composition," *Proc. 16th Int. Soc. Music Inf. Retr. Conf.*, pp. 591–597, 2015.

[3] M. M??ller, N. Jiang, and P. Grosche, "A Robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 3, pp. 531–543, 2013.

[4] J. C. Ross and P. Rao, "Detection Of Raga-Characteristic Phrases From Hindustani Classical Music Audio," *2nd CompMusic Work.*, pp. 133–138, 2012.

[5] V. Rao and P. Rao, "Vocal melody detection in the presence of pitched accompaniment using harmonic matching methods," *Proc. 11th Int. Conf. Digit. Audio Eff.*, pp. 1–8, 2008.

[6] O. Gotoh, "An improved algorith for matching biological sequences," *J. Mol. Biol.*, vol. 162 , pp. 705–708, 1982.