AUDIO METADATA EXTRACTION: THE CASE FOR HINDUSTANI CLASSICAL MUSIC

Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: prao@ee.iitb.ac.in

Abstract--**To make the vast digital archives of music more easily accessible, it is necessary to have searchable music descriptors, or metadata, that are meaningful and robust. While metadata conventionally covers factual information that accompanies the music on a CD such as genre, composer, artist, it could also include community-contributed semantic labels such as mood or other culture-specific tags. On the other hand, signal processing methods can be used to extract specific musical knowledge from audio signals such as descriptors related to the melody or rhythm which, in turn, depend to a great extent upon the particular music tradition. In this paper, we consider such acoustic metadata in the context of Hindustani classical music. Audio signal processing methods and data representations are discussed for specific retrieval tasks within the musicological basis of the tradition.**

I. INTRODUCTION

Musical metadata, or information about the music, is of growing importance in a rapidly expanding world of digital music consumption. To make the desired music easily accessible to the consumer, it is important to have meaningful and robust descriptions of music that are amenable to search. While music metadata conventionally covers editorial information that accompanies the music on a CD such as genre, composer and artist, it could also include community-contributed semantic labels such as mood or other culture-specific tags. MusicBrainz (musicbrainz.org) is an example of an online music information service with community contributed data that is a mixture of factual metadata as well as user tagging. An area of growing research importance is the automatic extraction of certain kinds of metadata from the audio signal. High-level descriptions such as genre, or melodic and rhythmic characteristics such as key, tempo and meter have been derived by using signal processing to obtain low-level acoustic attributes which are then incorporated in a machine-learning framework [1]. An important outcome of research on acoustic metadata extraction is the availability of computable measures of "music similarity". This leads to the possibility of expanding the scope of music search beyond the confines of textual tags. Pieces of music can be compared based on musical attributes obtained by audio content analysis, facilitating the discovery of new music by users.

Research in music information retrieval has been dominated by studies on Western music. With music being among the most culturally influenced types of content, it is becoming evident that new research is needed before music becomes universally accessible [2]. In this paper, we consider audio metadata in the context of Hindustani classical music. We review available work in audio signal processing and data representation within the musicological basis of the specific tradition. We consider, in particular, the extraction of melodic attributes for Hindustani vocal music, a prominent category of the genre. The next section gives an overview of the role of metadata in Hindustani music. This is followed by a presentation of audio signal processing methods and data representations for selected music retrieval tasks. We conclude with a brief overview of the potential of music computation for Indian classical traditions.

II. METADATA FOR HINDUSTANI MUSIC

Hindustani music is essentially solo music with a single main artist performing fixed compositions as well as improvising within a chosen melodic (*raga*) and rhythmic (*tala*) framework [3]. In a vocal concert, the singer is accompanied by a drone, tabla and sometimes the harmonium or sarangi. A concert is comprised of several stages such as the *alap*, *vistaar* and *taan* sections, each characterized by a particular performing style. Hindustani classical music collections are typically concert recordings commercially available as audio CDs. Editorial metadata provided on the CD cover is usually quite minimal and variable. While the artist name is always mentioned, additional information usually includes the sub-genre (e.g. *dhrupad*, *khyal*) name of the *raga* (melodic mode), *tala* (rhythmic cycle) and *laya* (tempo) of the various stages of the performance. In the case of vocal music, the title (*mukhda*) of the *bandish* (composition) may be mentioned as well. Concert recordings found on internet come with even less metadata, often just the artist name. It is therefore very attractive to consider the automatic extraction of music information from the audio signal.

The beauty and complexity of Indian classical music lies in the melody and rhythm (unlike Western music where harmony is emphasized). Thus *raga* and *tala* information form the most essential descriptions and the bases for most music searches. *Raga* identity is cued by the permitted pitch scale intervals (*swara*) and their hierarchy, commonly used melodic phrases (*pakad*) and or-

namentation. In principle, audio processing can be applied to achieve a full transcription of the audio that would be useful for retrieval based on melody or rhythm, as well as in musicology research and pedagogy, considering especially the absence of written scores in this oral tradition. In the next section, we discuss the audio processing challenges and present some methods that address these.

III. Signal Processing

The melody of a piece of music is related to the time-varying pitch of the predominant voice. In the case of vocal music, this is the "tune" held by the singer. A melody is represented by a sequence of notes each specified by its pitch and duration. In the case of Indian classical music, a discrete-pitch note representation is inadequate due the importance of ornamentation (various shapes of pitch variation in time) and specific context dependent intonation of *swaras* (scale intervals). The rhythm is specified by the *tala* which relates to the arrangement of tabla strokes in a rhythmic cycle. The vocal pitch and tabla stroke onsets are low-level acoustic features, relating to melody and rhythm respectively, that can be detected by suitable signal processing methods as discussed next.

A. 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 (i.e. fundamental frequency, F0) of the predominant source in the polyphonic mix. Melody detection involves identifying the vocal segments and tracking the pitch of the vocalist. The tanpura and harmonium are strongly pitched instruments. A conventional monophonic pitch detector based on assumptions of a single harmonic source is unsuitable, and it is necessary to extract the instantaneous F0's of all the concurrent pitched sources as an intermediate stage. Next, pitch saliency and continuity constraints can be applied to estimate the predominant pitch corresponding to the melodic voice. An example of such a pitch detector specifically developed for tracking the singing voice uses sinusoids detected by short-time spectral analysis to identify the multiple pitches present locally. These are subsequently pruned based on estimated harmonic source strength and continuity in time [4]. On occasion when an accompanying instrument such as the harmonium is relatively loud, the system tracks two pitch contours simultaneously and uses the difference in temporal characteristics (steady pitches of the harmonium versus the more continuously varying pitches of the singing voice) to select the vocal melody. The same characteristics are used to eliminate the purely instrumental regions from the tracked pitch contour [5].

Apart from interference from pitched accompaniment, pitch detection is complicated by the nature of pitch variation. Hindustani vocal music is characterized by precisely intoned steady notes as well as rapid pitch modulations including transitions between notes and ornamentation. Often specific ornaments are characteristic of a particular raga hinting at the importance of accurate pitch tracking in melody-based retrieval tasks. In regions of rapid pitch variation, the harmonic components are highly non-stationary and the accuracy of the initial short-time spectral analysis depends critically on the window duration. While longer windows improve the frequency resolution of slowly varying harmonic components, regions of rapid pitch variation require analysis windows to be short enough for the accurate estimation of instantaneous frequencies and amplitudes. Automatically adapting the window length to the underlying signal characteristics may be achieved efficiently by the maximization of a signal sparsity measure computed from the short-time spectrum at each analysis instance [6]. A "sparse" short-time spectrum is expected to have more concentrated components and hence provides for more accurate detection of sinusoidal parameters. The normal analysis window duration of 40 ms is reduced to 30 or 20 ms depending on the rate of variation of the signal components within the window.

Fig. 1 shows the spectrogram of an audio segment extracted from a concert of famous *khyal* vocalist Kishori Amonkar in the *raga* Deskar, *tala* Tintal (16 beat) and Vilambit (slow) *laya*. The 30 sec long clip is a section of the *bol-alap* where she improvises in the *bandish* Piya Jaag. We see the distinct vocal regions marked by the strong and rapidly varying voice harmonics in a background of tanpura, harmonium and swarmandal partials that are relatively stable. We also observe the strong low frequency components of the tabla in the region below 100 Hz. These arise from the tabla strokes (thin vertical lines) and decay rapidly. Superimposed on the spectrogram is the detected vocal pitch contour in white. We observe that it captures the time-varying vocal pitch accurately.

B. Rhythm detection

The *tala* is the rhythm structure comprising of equal duration beats divided into subgroups. The exact structure of a cycle is represented by the sequence of tabla *bols*. However detecting and identifying individual *bols* in the signal mixture of voice and instruments is challenging. Identifying prominent periodicities in the sequence of strokes is more practical and can reveal the musical meter (e.g. duple, triple) and the tempo of the piece [7].

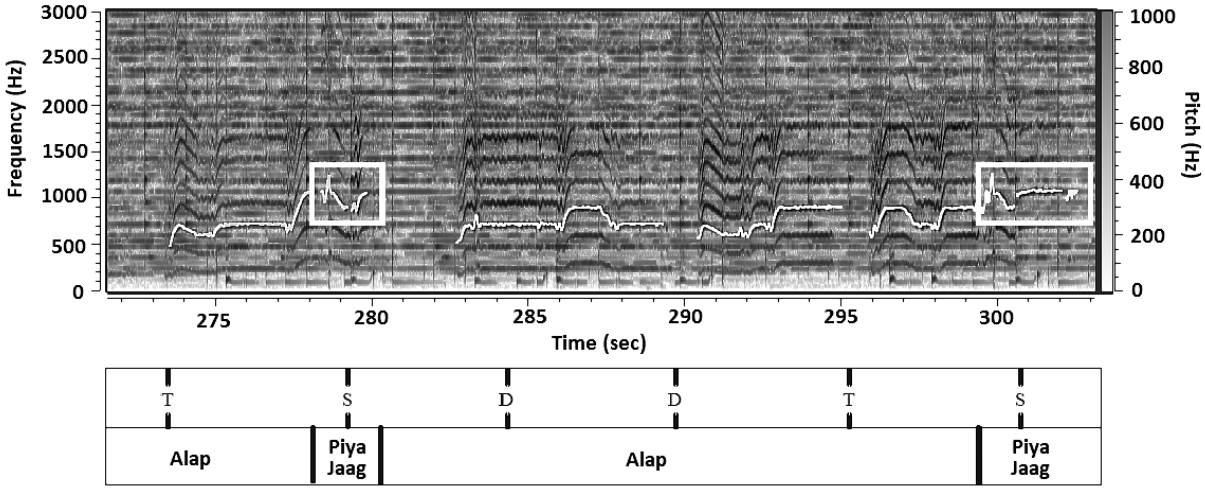

Figure 1. Top: spectrogram of an audio segment of "Piya Jaag" by vocalist Kishori Amonkar with superposed vocal pitch contour (the *mukhda* is in boxes); below: first beats of each sub-cycle of Tintal (S= *sam*) with aligned lyrics.

As mentioned earlier, in Fig. 1 the onsets of tabla strokes are visible as high energy vertical striations in the spectrogram. The onsets are most prominent in the higher frequency region due to the absence of strong vocal components here. The audio signal is filtered to retain the band [5000, 8000] Hz. The filter output power is subjected to a first-order difference and then half-wave rectified. The detected onsets correspond to the tabla strokes comprising the 16-beat Tintal cycle. The first beat (*sam*) is labeled "S". The remaining labels in Fig. 1 correspond to the first beat of each 4-beat sub-cycle. We see that the duration of the cycle is over 20 seconds, consistent with *vilambit laya* (slow tempo).

IV. Melody Based Retrieval

In the previous section, we considered the computation of the melodic contour from the audio signal. The continuous contour, an example of which appears in Fig. 1, captures completely the melodic dimension of the music piece. Thus information about the high level attributes of the music such as the scale or melodic mode (*raga*), characteristic phrases and the specific tune as well, are all contained in the melodic contour. In order to exploit this information for retrieval tasks, we need a suitable data representation that can be used in the appropriate similarity matching framework. As mentioned in Sec. 3, melody detection involves extracting the fundamental frequency (F0), measured in Hz, of the vocal source as it varies with time. The subjective perception of pitch however is related to the log of F0 so that a constant pitch change refers to a constant *ratio* of fundamental frequencies. Thus a musically relevant representation of the pitch time series is with respect to the logarithmic frequency scale where an octave is divided into 1200 equal intervals (known as cents).

In Western music, with its emphasis on discrete pitches and absolute tuning and relatively limited forms of ornamentation, the continuous pitch contour corresponds closely to a sequence of stable notes with discrete pitch values and durations. The discrete pitches constitute the equitempered scale with 12 semitones (100 cents apart) per octave. The musical score in symbolic notation is thus a near equivalent of the melodic contour and music retrieval tasks can operate at the level of symbolic string matching. In Indian classical music, on the other hand, symbolic notation proves inadequate to deal with tuning variations and complex ornamentation that are fundamentally linked to *raga* characteristics. Being an oral tradition, this aspect has not seriously hampered music education. However for the music retrieval task, there is a need for more complete data representations to be derived from the continuous pitch contour. In this section, we present three examples that illustrate the problem of data representation and similarity modeling for specific melodic matching tasks.

A. *Motif Identification*

An important section of the Hindustani *khyal* vocal concert centers around improvisation embedded within a raga-specific composition known as *bandish*. The singer elaborates within the raga's framework in each rhythmic cycle before returning to the main phrase of the bandish (*mukhda*) which acts like the refrain. The automatic detection of this repetitive phrase, or motif, from the audio signal would contribute to important metadata concerning the identity of the *bandish*. The *mukhda* is characterized by its melodic shape (and also its lyrics). A suitable representation of the melodic shape and a matching criterion can help to find music corresponding to a desired *bandish*.

Fig. 2. shows a few instances of manually labeled pitch contour segments (approximate duration 4 sec) corres-

ponding to the *mukhda* of "Piya Jaag", an audio segment of which recording appears in Fig. 1. The horizontal lines show the *swara* locations as determined from the *raga* and the tonic of the singer. The vertical lines indicate the beat locations in the 16 beat rhythmic cycle. We observe that the *mukhda* phrases are similarly aligned with respect to the rhythmic cycle. The onset of the syllable "Jaag" coincides with the *sam*, a characteristic that is strictly maintained providing certain bounds on the improvisation within a cycle. While the *mukhda* phrases correspond in theory with the note sequence [Da, Pa, Ga, Pa], we observe a high degree of variation in the attained melodic shapes across the different instances. The variability lies in both, relative durations of the syllables and the actual pitch values traversed between note locations. We seek a distance measure that takes a low value between the time series of pitch values corresponding to two *mukhda* phrases while clearly discriminating between *mukhda* and non-*mukhda* segments. A dynamic time-warping (DTW) based distance measure applied to the pitch time segments with local cost linked to pitch difference in cents above a threshold value (to account for imperceptible differences) shows promising results [8]. To reduce the search space, only segments that are similarly aligned with the underlying rhythmic structure are selected for distance computation with respect to the reference *mukhda* segment. The same method can be extended to the detection from audio of other phrases characteristic of a particular composition or *raga*.

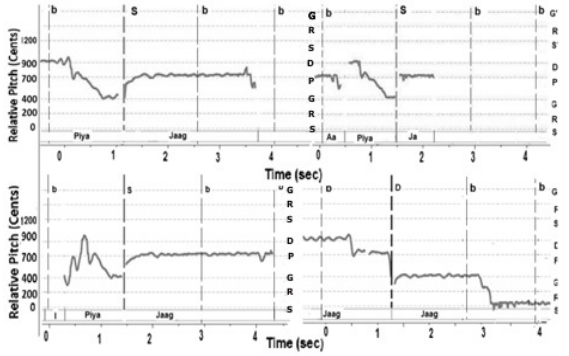

Figure 2. Pitch contour segments of three instances of mukhda 'Piya Jaag' from the same concert. Bottom, left is a different phrase.

B. Ornament Verification

Apart from phrases, comprising a sequence of notes realized by a continuous pitch contour, ornaments constitute another important component of the melody. Ornaments are melodic shapes of relatively brief durations (within 0.5 sec or so), and range from simple glides between steady notes (*meend*) and oscillations (*gamak*) to a combination of the two, and more. Ornaments are an essential component of raga identity, and the correct rendition of an ornament is the mark of a well trained singer. It is useful to consider how the data representation and similarity matching methods of music retrieval may be applied in evaluation tools for music learning. In this application, we would be interested in evaluating the perceived similarity of a learner's rendition of an ornament to a reference version of the ornament.

Fig. 3 shows the pitch contour of an extracted audio segment corresponding to a particular transition between notes. The reference contour is obtained from a classical song rendered by famous playback singer Asha Bhonsle while the remaining contours are extracted from recordings of the same song by singers of different proficiency levels. The transition by the reference singer is achieved by an overall down-glide with superimposed oscillations where each oscillation corresponds to a note. While a trained musician would easily name the pitch intervals (*swara*) realized by the melodic contour, it is known that the precise interval (measured by either the mean or the range of an oscillation) varies widely from instance to instance within and across artists rendering the same composition. Through several pertinent examples from Carnatic music practice, Subramanian [9] demonstrates that the similarity lies in the overall melodic shape. Subjective listening tests involving musicians indicated that perceived similarity between the reference and a test segment is related to certain salient properties of the melodic shape such as the duration of the down-glide and the rate and amplitude of the oscillation [10]. The down-glide is characterized by a 3rd degree polynomial fit (shown as dashed curves in Fig. 3) and the oscillations, by the frequency and amplitude. The frequency of oscillation is more critical than the amplitude. A simple decision-tree classifier using the three shape features was successfully trained on subjectively labeled (good/medium/bad) sets of test singers' ornaments [10].

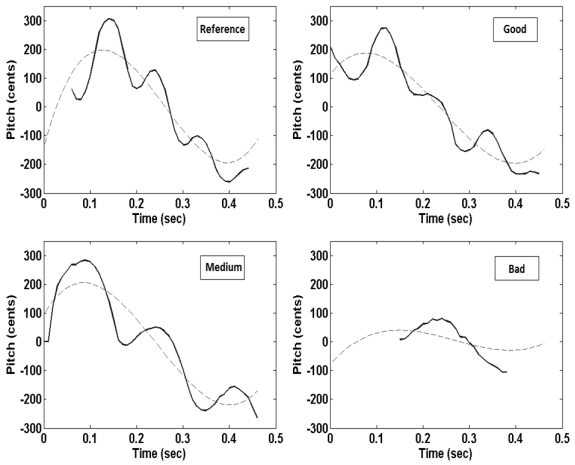

Figure 3. A reference ornament pitch contour (top, left) rendered by singers of various proficiencies.

C. Raga Classification

We consider the problem of detecting *raga* identity from a segment of a Hindustani classical music recording. A *raga* is not a fixed composition but rather a melodic framework. One important (even if insufficient) distinguishing attribute of a *raga* is the tone material (i.e. allowed *swara* or scale intervals) and their hierarchy. In

view of this, pitch distributions have been previously applied in *raga* classification [11]. Thus the data representation is the first-order distribution of pitch intervals (in cents with respect to the tonic) rather than a pitch sequence. Unlike Western music where a pitch-class distribution over the 12 semitones of the octave describes the scale of the music, we require a relatively fine resolution of the pitch axis for considerations similar to those observed for Turkish makam music [12].

Fig. 4 shows pitch histograms using 10 cent bin widths (i.e. 120 divisions of the octave) for two ragas, Marva and Puriya, performed by Pt. Vidhyadhar Vyas [3]. Marva and Puriya have the same set of *swaras*. From the theory, Marva has Re (r) and Dha (D) as its stressed notes, while Ga (G) and Ni (N) are stressed in Puriya. We observe that the pitch histograms clearly capture these differentiating aspects. Further, it is held that Re and Dha are intoned slightly low in Marva [3]. It is interesting to observe this coming out in the pitch histogram as well. Pitch histograms have been previously applied to validate musicological observations on relative intonation (shrutis) of specific raga pairs [13]. The KL distance between pitch distributions was used to implement classification of test recordings within a raga pair.

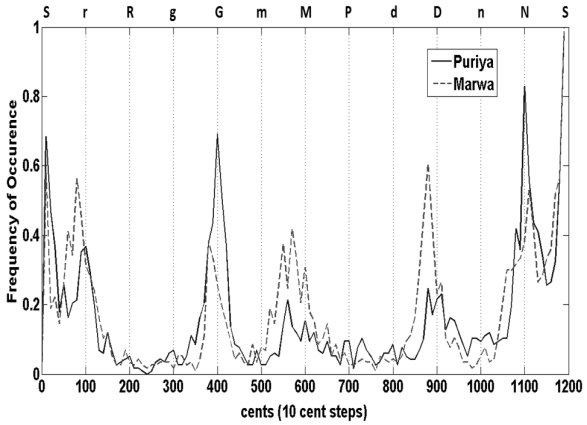

Figure 4. Pitch interval histograms from Marva and Puriya recordings. The vertical lines indicate equitempered intervals.

V. CONCLUSION

Metadata, crucial to improving access to the world's music, is observed to be culture-specific, both in terms of characteristics of the particular genre or tradition as well as user expectations. The automatic extraction of high-level attributes related to the musical dimensions of melody and rhythm has been considered for the case of Hindustani classical music. Specific data representations and distance measures have been discussed in the context of melodic similarity based tasks. The scope of the study can be extended to the discovery from audio of characteristic phrases and thus the incorporation of more complete knowledge in *raga* recognition [14]. Computational descriptions of *tala* are yet to be extensively explored. Other musical dimensions such as timbral and loudness dynamics play an important role in phrase intonation which computational methods could potentially help uncover. The discrimination of the different sub-genres and styles (*gharanas*) constitutes yet another interesting challenge.

Music computing research holds a vast potential for Indian classical traditions, with the surface barely scratched so far in spite of firmly grounded musicological, cultural and social connections. While retrieval constitutes the most obvious beneficiary of automatic metadata extraction tools, rich transcriptions of the audio in terms of musically relevant segmentation can enhance the music listening experience greatly. Music education and musicological studies too stand to benefit from the tools developed for music retrieval.

REFERENCES

[1] M. Casey, R. Veltekamp, M. Goto, M. Leman, C. Rhodes and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. of the IEEE*, vol. 96, no. 4, 2008.

[2] X. Serra, "A multicultural approach in music information research," *Proc. of ISMIR* 2011, Miami.

[3] J. Bor, Editor, S.Rao, W. van der Meer, J. Harvey, Authors, *The Raga Guide: A Survey of 74 Hindustani Ragas*, Nimbus Records with the Rotterdam Conservatory of Music,1999.

[4] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 8, 2010.

[5] V. Rao, C. Gupta and P. Rao, "Context-aware features for singing voice detection in polyphonic music", *Proc. of Adaptive Multimedia Retrieval* 2011, Barcelona, Spain.

[6] V. Rao, P. Gaddipati and P. Rao, "Signal-driven window-length adaptation for sinusoid detection in polyphonic music," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 20, no.1, 2012.

[7] S. Gulati and P. Rao, "Meter detection from audio for Indian music," *Proc. of FRSM-CMMR* 2011, Springer LNCS, Editors: S. Ystad et al.

[8] J. Cheri Ross, Vinutha T.P. and P.Rao, "Detecting melodic motifs from audio for Hindustani classical music," *Proc. of ISMIR* 2012.

[9] Subramanian, M., "Carnatic RagamThodi – Pitch analysis of notes and gamakams," *Journal of the Sangeet Natak Akademi*, XLI(1), 2007.

[10] C. Gupta and P. Rao, "An objective evaluation tool for ornamentation in singing," *Proc. of FRSM-CMMR* 2011, Springer LNCS, Editors: S. Ystad et al.

[11] P. Chordia and A. Rae, "Automatic raag classification using pitch-class and pitch-class dyad distributions," *Proc. of ISMIR* 2007.

[12] A.C. Gedik and B. Bozkurt, "Pitch-frequency histogram-based music information retrieval for Turkish music," *Signal Processing,* vol. 90, 2010.

[13] S. Belle, R. Joshi, and P. Rao, "Raga Identification by using swara intonation," *Journal of ITC Sangeet Research Academy*, vol. 23, 2009.

[14] 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. XVII (3&4), 1989.