QUANTIZED MELODIC CONTOURS IN INDIAN ART MUSIC PERCEPTION: APPLICATION TO TRANSCRIPTION

Ranjani, H. G.
Dept of ECE,
Indian Institute of Science,
Bangalore
ranjani@iisc.ac.in

Deepak Paramashivan
Dept of Music,
University of Alberta,
Canada
paramash@ualberta.ca

Thippur V. Sreenivas
Dept of ECE,
Indian Institute of Science,
Bangalore
tvsree@iisc.ac.in

ABSTRACT

Rāgas in Indian Art Music have a florid dynamism associated with them. Owing to their inherent structural intricacies, the endeavor of mapping melodic contours to musical notation becomes cumbersome. We explore the potential of mapping, through quantization of melodic contours and listening test of synthesized music, to capture the nuances of *rāgas*. We address both Hindustani and Carnatic music forms of Indian Art Music. Two quantization schemes are examined using stochastic models of melodic pitch. We attempt to quantify the salience of *rāga* perception from reconstructed melodic contours. Perception experiments verify that much of the *rāga* nuances inclusive of the *gamaka* (subtle ornamentation) structures can be retained by sampling and quantizing critical points of melodic contours. Further, we show application of this result to automatically transcribe melody of Indian Art Music.

1. INTRODUCTION

Melody contours are often perceived as continuous functions though generated from notes which assume discrete pitch values. The rendition of a *rāga*, the melodic framework of Indian Art Music (IAM), is a florid movement across notes, embellished with suitable ornamentations (*gamakas*). Several engineering approaches to analyse and/or model pitch contours rely on 'stable' notes [5, 12]; yet, it contradicts the perceptions and claims of musicians in both Carnatic and Hindustani forms of music and also that of detailed experiments which assert that it is actually the manner of approaching notes that characterizes a *rāga* [1, 6]. Algorithms to automatically align note transcription to melodic contours show promise more at a rhythm cycle level rather than at a note level [21], leading to a hypothesis that it is necessary to study the role of pitch in rendering notes rather than finding / transcribing notes from pitch contours [11]. The difficulty involved in identifying notes from a rendered melodic contour can be seen in Figure 1.

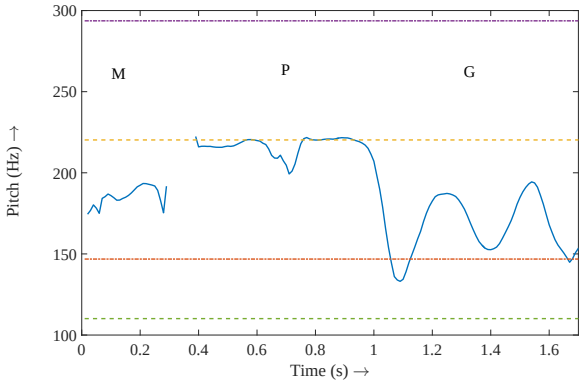

Figure 1. [Color Online] Contour of a vocal melodic clip (rendition by Vid. Deepak Paramashivan) in *Thodi rāga* of Carnatic music form (tonic frequency at 146.83 Hz). The transcribed notes correspond to '$MaPaGa$'. The presence of ornamentations pose difficulty for transcription.

In this work, we analyze pitch contours in an attempt to understand (i) how musicians possibly assess a correctly rendered note (ii) how they approach subsequent note(s) in a *rāga*. We explore the possibility to incorporate this understanding to engineer an automated framework to represent a *rāga* in terms of note sequences. A perceptual study of the effects of two quantization schemes on *rāga* characteristics (*rāga-bhava*) is explored. For a more detailed exposition of *rāgas* in Indian Art Music, interested readers can refer to [19, 20].

1.1 Complexity of Pitch Contours in Indian Art Music

A *rāga* contains 3 structures of information : (i) Pitch positions of notes (*swarasthāna*) (ii) Ornamentation of notes (*Gamaka*) (iii) Note movement (*swarasanchāra*). All the three structures are coupled in a *rāga* rendition. The note position is embellished with *gamakas*, and is also dependent on the note transitions themselves.

In [13], different notes and their transitions are studied and classified as 'inflected intervals', 'transient notes' and 'transient inflexions', while acknowledging that musi-

cal insight and knowledge is necessary to distinguish between different transitions.

From an engineering perspective, in a *rāga*, we observe that inspite of all such note transitions detected at both finer and coarser levels, the peaks, valleys and stable regions correspond to discrete pitch values with an error factor. It is also logical that a musician perceives these points to assess if the intended note has been reached [1] . The salience of peaks, valleys and stable regions is utilized for motif spotting in Carnatic music in [10]. However, not all peaks, valley regions can correspond to notes as per conventional transcriptions [15]. As an engineering approach, we propose to discretize or quantize the pitch contours at these critical points using a semi-continuous Gaussian mixture model (SC-GMM) proposed in [18] (Section 2) and thus map continuous pitch contours to discrete note sequences. We believe such a mapping brings us closer to understanding the structures of *rāgas* in accordance with the theory that discrete elements carry the structural information of musical forms while expressions are realized in continuous variations [14].

While analysing contours w.r.t. discrete pitch values, we often encounter scenarios in which pitch values can either overshoot or undershoot the intended values as can be seen in Figure 1; this is also reported in literature [13, 15]. The following reasons can be attributed to such detours w.r.t. discrete pitch values - (i) Performers' intent to generate certain perceived effect in the listener (ii) Possible deviations during learning/ fast renditions (iii) Creative freedom and margin of error allowed in rendering a *rāga* as an art form. Any deviation which does not bring about the required perceptual effect can cause a connoisseur/musician to not appreciate the rendition in its totality.

In this work, we assume the deviations to be due to any of above reasons and hence is part of errors in quantizing pitch contours. If the quantization process has disregarded the musically intended overshooting and undershooting of pitch values, it only implies that the effect of the *rāga* is not captured completely in the quantized sample. In order to analyze the importance of limits of quantization, we reconstruct the melody from quantized sequences and conduct perception experiments on these melodies (Section 3.3). Further, we propose a framework by using these quantized notes to transcribe a contour (Section 4.1).

2. QUANTIZATION MODEL

Given pitch contours, $y(t)$ estimated from audio recordings, it is possible to identify the tonic frequency f_T as shown in [7, 18]. The pitch contours are tonic normalized and mapped to a common tonic, f_U; let $y_n(t) = y(t) * f_U/f_T$ denote pitch contours mapped to common tonic frequency [2] . This helps to analyze different renditions of same *rāga*. Let $\tau = \{\, t \mid \nabla y_n(t) = 0 \,\}$ be the set of critical points and $\boldsymbol{x} = \{\, y_n(\tau) \,\}$ be the corresponding critical pitch values. The tuple $\mathbf{X} = (\boldsymbol{x}, \tau)$ are the critical points of $y_n(t)$. Mathematically, critical points can be obtained only if a function is differentiable. We estimate $\mathbf{X}$ from the zero crossings of numerical gradient of $y_n(t)$.

2.1 Semi-Continuous Gaussian Mixture Model

Consider the Semi-Continuous Gaussian Mixture Model (SC-GMM) [18] with K number of components whose means, $\mu_k \forall k \in \{1, 2, \ldots, K\}$ within an octave are fixed in accordance to the note ratios used in IAM, as shown in Table 1. The distribution of pitch values in y_n and the critical pitch values $\boldsymbol{x}$ can be modeled using SC-GMM as:

$$p(y) = \sum_{k=1}^{K} \alpha_{k,y} \mathcal{N}(y_n; \mu_k, \sigma_{k,y}) \quad (1)$$

$$p(x) = \sum_{k=1}^{K} \alpha_{k,x} \mathcal{N}(x; \mu_k, \sigma_{k,x}) \quad (2)$$

For a fixed K components, the set of parameters estimated from distribution of pitch are $\{\boldsymbol{\alpha}_y, \boldsymbol{\sigma}_y\}$ and $\{\boldsymbol{\alpha}_x, \boldsymbol{\sigma}_x\}$. $\boldsymbol{\mu}$ parameters are not estimated since they are fixed and are same in both cases.

2.2 Quantization using SC-GMM

We use the above model to quantize pitch contours. Each pitch sample of $y_n(t)$ can be quantized to a nearest component of SC-GMM which maximizes its probability:

$$k_y^*(t) = \underset{k \in \{1,2,\ldots,K\}}{\arg\max}\ \alpha_{k,y} \mathcal{N}(y_n(t); \mu_k, \sigma_{k,y}) \quad (3)$$

Similarly, every critical pitch of $\boldsymbol{x}$ can be quantized as:

$$k_x^*(\tau) = \underset{k \in \{1,2,\ldots,K\}}{\arg\max}\ \alpha_{k,x} \mathcal{N}(x(\tau); \mu_k, \sigma_{k,x}) \quad (4)$$

Thus, both y_n and $\boldsymbol{x}$ are now quantized and correspond to a sequence of notes; their temporal information (corresponding to $\{t\}$ and $\{\tau\}$) are retained.

3. SYNTHESIS FROM QUANTIZED SEQUENCE OF NOTES

To check if this mapping process captures the essence of *rāgas* and to assess the effect of quantization on *rāga* perception, we conduct perception experiments. Audio clips are synthesized for perception. In order to synthesize audio from quantized note sequences, we first synthesize melody contours, and use the same to synthesize audio clips.

3.1 Quantized Pitch Contour

The un-quantized $y_n(t)$, the quantized $k_y^*(t)$ and $k_x^*(\tau)$ are interpolated to obtain a contour sampled at F_s, the sampling frequency of the discrete-time audio signal [3] . Piecewise cubic hermite interpolating polynomial is used with

[1] This also explains the fact that music listeners do not perceive intermediate notes during note transitions which are greater than a semitone; for example, when a musician glides from Sa (tonic) to Pa (fifth) in a *rāga*, we do not perceive all the intermediatary semi-tones which the glide passes through.

[2] In this work, f_U is chosen as 146.83 Hz corresponding to D_3 note of Western scale.

[3] $y_n(t)$ also requires interpolation as it is estimated at frame rate coarser than F_s.

$(t, y_n(t))$, $(t, \mu_{k_y^*}(t))$ and $(\tau, \mu_{k_x^*}(\tau))$ being knots for each of the interpolation. These result in 3 different pitch contours for signal synthesis. The pitch contour obtained from interpolating $(t, y_n(t))$ can be considered (for all practical purposes) as a reference pitch contour. A comparison of pitch contours obtained by interpolating (τ, k_x^*) and $(t, k_y^*(t))$ are shown in Figure 2.

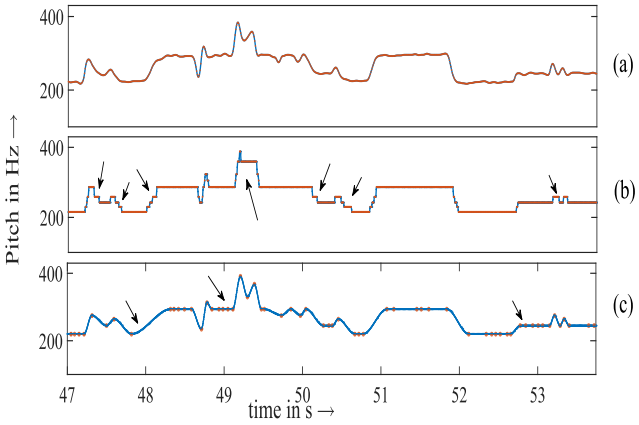

Figure 2. [Color Online] Pitch contours obtained by interpolation of (a) $y_n(t)$ (b) $k_y^*(t)$ and (c) $k_x^*(\tau)$ (*Sāveri rāga*). $f_U = 146.83\ Hz$ for all 3 contours. Red dots are knots of interpolation; solid blue lines are the interpolated contours. Arrows highlight local distortions in interpolated contours.

We see from Figure 2(b) that quantizing every pitch sample $(k_y^*(t))$ can potentially lead to introducing additional stable notes (which are not perceived in the original contour). This may result in an unfavorable shape of the contour and hence dynamics of notes may be altered during interpolation. The advantage of quantized critical pitch samples $(k_x^*(\tau))$ is that, on interpolation, the dynamics around a note are more likely to be retained as seen in Figure 2(c). Some subtle *gamakas*, in-spite of being captured, could be quantized onto single note due to statistical weight of an adjacent note, resulting in synthesizing a flat region. In this work, pitch values at critical points are perturbed while the critical points themselves are unaffected; hence slopes might be perturbed in a different manner in various sections of the pitch contours and hence perceptual effect of the same is not simple to predict.

3.2 Synthesizing Audio

The audio signal for perception can be synthesized from the interpolated pitch contours (sampled at F_s) using the time-varying sinusoidal synthesis model. The model can be expressed as:

$$\hat{s}_f(t) = a(t) * \left(\sum_{i=1}^{H} sin\left(\frac{2\pi h}{F_s} \int_0^t f(t)dt \right) \right) \quad (5)$$

where $a(t)$ represents the vocal-tract shaping filter, $*$ is the convolution operation, H denotes the number of harmonics, F_s is the sampling frequency, $f(t)$ represents the pitch contour which is to be synthesized (explained in Section 3.1) and $\int_0^t$ is approximated as cumulative sum for discrete implementation. The vocal-tract shaping filter is chosen to be time-invariant and is that of vowel /ā/. An all-pole model is used to synthesize the transfer function using formant frequencies and bandwidth of /ā/ as $(730, 1090, 2440, 3781, 4200)\ Hz$ and $(60, 50, 102, 309.34, 368)\ Hz$ respectively [17]. A drone signal is added to the synthesized audio so that reference tonic is present in it.

3.3 Perception Test Experiments

Let $\hat{s}_{ref}(t)$ be the audio signal synthesized from interpolated $(t, y_n(t))$, $\hat{s}_y(t)$ be the audio synthesized from interpolated $(t, k_y^*(t))$ and $\hat{s}_x(t)$ synthesized from interpolated $(\tau, k_x^*(t))$. We quantize using both 22-note and 12-note intervals to study the effect of number of quantization levels on *rāga* perception i.e., $\hat{s}_{y_{22}}(t)$ and $\hat{s}_{x_{22}}(t)$ are the audio signals synthesized using (3) and (4) with $K = 22 * 3$ (covering 3 octaves), while $\hat{s}_{y_{12}}(t)$ and $\hat{s}_{x_{12}}(t)$ correspond to $K = 12 * 3$ levels. The means within an octave of the SC-GMM are as chosen according to Table 1.

We choose certain *rāgas* along with the corresponding pitch features from the publicly available Carnatic and Hindustani music database used in [8, 9]; in this database, pitch has been estimated every $4.44\ ms$ using Essentia [3].

3.3.1 Comparison of $\hat{s}_y(t)$ and $\hat{s}_x(t)$

As argued earlier, we hypothesize $\hat{s}_x(t)$ to be a closer representative of $\hat{s}_{ref}(t)$ than $\hat{s}_y(t)$. To verify which among $\hat{s}_x(t)$ is indeed perceptually closer to $\hat{s}_{ref}(t)$, a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) kind of experiment is performed. We select $K = 12 * 3$ for quantization levels. 6 musically trained listeners are tasked with three experiments - different reference clips (average 7 s duration) from *Nattai rāga* of Carnatic music are presented in each experiment. Within each experiment - (i) $\hat{s}_{y_{12}}(t)$ (ii) $\hat{s}_{x_{12}}(t)$ (iii) hidden $\hat{s}_{ref}(t)$ - form 3 audio stimuli presented to listeners in randomized order along with an explicit reference clip $\hat{s}_{ref}(t)$. Listeners are asked to rate the closeness of each to the reference signal on a scale of 1-100 (100 implies the stimuli is indistinguishable from the reference). We refer to this as perception test 1 (PT-1).

From the results, $\hat{s}_{y_{12}}(t)$ is consistently rated least by all the listeners. These audio clips are perceived to be 'electronic', with temporal distortion clearly heard. Listeners have rated $\hat{s}_{x_{12}}(t)$ at an average of 88.88% close to explicit reference, while the hidden reference $\hat{s}_{ref}(t)$ is rated at an average of 96.5% closeness to explicit $\hat{s}_{ref}(t)$; this is because $\hat{s}_{x_{12}}(t)$ is confused with the hidden $\hat{s}_{ref}(t)$ in 38.8% cases by the subjects. $\hat{s}_{y_{12}}(t)$ is rated at an average 56.27% close to explicit reference. The bane of synthesizing melody with $k_y^*(t)$ sequence is easily perceivable by all trained listeners. This validates our hypothesis that $\hat{s}_x(t)$ is closer to $\hat{s}_{ref}(t)$ than $\hat{s}_y(t)$.

3.4 *Rāga* Perception Experiment

With $\hat{s}_x(t)$ being a close model of $\hat{s}_{ref}(t)$, we hypothesize much of *gamaka* structures in a *rāga* rendition is retained; also, micro-tonal dynamics of *rāga* might be better captured with $K = 22 * 3$ levels than $K = 12 * 3$.

Note name	22 Notes (Position)	Pitch Ratio	12 Notes(Position)
Sa	S (1)	**1**	S(1)
Ri	R_{11} (2)	256/243	
	R_{12} (3)	**16/15**	R_1(2)
	R_{21} (4)	10/9	
	R_{22} (5)	**9/8**	R_2/G_1(3)
Ga	G_{11} (6)	32/27	
	G_{12} (7)	**6/5**	R_3/G_2(4)
	G_{21} (8)	**5/4**	G_3(5)
	G_{22} (9)	81/64	
Ma	M_{11} (10)	**4/3**	M_1(6)
	M_{12} (11)	27/20	
	M_{21} (12)	**45/32**	M_2(7)
	M_{22} (13)	729/512	
Pa	P (14)	**3/2**	P(8)
Da	D_{11} (15)	128/81	
	D_{12} (16)	**8/5**	D_1(9)
	D_{21} (17)	5/3	
	D_{22} (18)	**27/16**	D_2/N_1(10)
Ni	N_{11} (19)	16/9	
	N_{12} (20)	**9/5**	D_3/N_2(11)
	N_{21} (21)	**15/8**	N_3(12)
	N_{22} (22)	243/128	
Sa'	S(next octave)(23)	**2/1**	S (upper octave)(13)

Table 1. Pitch ratios in an octave for 22-note system of Indian Art Music. The ratios used in 12-note system are in bold face.

3.4.1 Experimental Setup

We choose some *rāgas* (shown in Table 2) which are considered by experts to be musically challenging to render as they contain lot of *gamakas* and micro-tonal structures. For each listener, two different renditions (by different singers) are presented for every *rāga*. The singer identity is masked as a result of time-invariant /ā/, the shaping filter for the pitch contour; hence any bias factor due to singer in the listening experiments is reduced.

	Carnatic music	Hindustani music
1.	Begada	Bhairav
2.	Bhairavi	Darbari
3.	Saveri	Marwa
4.	Sahana	Puriya Dhanashree
5.	Sindhu Bhairavi	Yaman
6.	Thodi	

Table 2. *Rāgas* chosen for perception experiment.

To verify if $\hat{s}_x(t)$ captures the *rāga* nuances along with the *gamakas* in its entirety as represented in $\hat{s}_{ref}(t)$, in each experiment, we present a 1 min duration clip of $\hat{s}_{ref}(t)$ and its corresponding (i) $\hat{s}_{x_{22}}(t)$ and (ii) $\hat{s}_{x_{12}}(t)$ (synthesized) audio clips [4].

We first present to music experts, $\hat{s}_{ref}(t)$ as the reference and ask them to rate on a scale of 1-10 for *rāga* characteristics present in $\hat{s}_{ref}(t)$. The same listener is now presented with $\hat{s}_{x_{22}}(t)$ and $\hat{s}_{x_{12}}(t)$ (not necessarily in that order) and asked to rate closeness of each with respect to *rāga* nuances of $\hat{s}_{ref}(t)$ on the scale of 1-10. Lower rating implies *rāga* nuances are compromised due to quantization. Thus, each listener for Hindustani music form participates in 10 (5 *rāgas* with 2 different renditions) such experiments; and, 12 experiments are presented for each Carnatic expert listener. This is perception test 2 (PT-2).

3.4.2 PT-2 Results and Analysis

5 performing Carnatic musicians were selected for the perception test in Carnatic music; similarly, 5 musicians trained in Hindustani music were considered for the Hindustani music perception tests.

The average ratings of $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$ w.r.t. reference $\hat{s}_y(t)$ for each *rāga* considered in Carnatic and Hindustani music is as shown in Figure 3 (a) and (b) respectively.

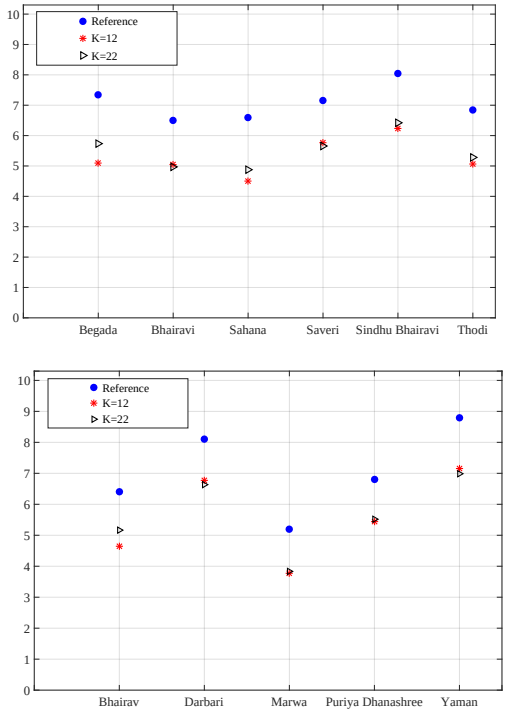

Figure 3. [Color Online] Perception rating for *rāga* characteristics for $\hat{s}_y(t)$, $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$ for (a) Carnatic music averaged over 5 listeners (b) Hindustani music averaged over 5 listeners.

The $\hat{s}_{ref}(t)$ ratings absorbs anomalies such as sudden breaks and octave errors which commonly occur in pitch

[4] The clip is a part of the starting portion of the original rendition but between the region 30 s to 90 s. While the *rāga* characterizing phrases will be brought about initially, we hypothesize that *rāga* nuances must be showcased at any chunk of time.

estimation algorithms, as well as errors committed by the artist in the rendition. In both Carnatic and Hindustani music clips, $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$ are reported to be quite close to $\hat{s}_{ref}(t)$ and requires more careful inspection by repeated contrasts against $\hat{s}_{ref}(t)$ as described ahead.

In order to pin-point distortions perceived, expert listeners had to intently re-listen the $\hat{s}_{ref}(t)$ multiple times to confirm if perceived foreign/distorted notes are present in $\hat{s}_{x_{12}}(t)$ or $\hat{s}_{x_{22}}(t)$ clips and not in $\hat{s}_{ref}(t)$ clip itself. Some of the commonly observed distortions are: *gamakas* being flattened, a foreign note perceived in $\hat{s}_{ref}(t)$ clip being corrected and note shift at micro-tonal level. A Hindustani music performer after multiple listenings, could report ~ 13 perceivable distortions (in each $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$) with respect to total 10 reference clips.

In the perception experiment in Hindustani music, at $K = 12 * 3$ and $K = 22 * 3$ quantization levels, given $\hat{s}_{ref}(t)$, expert listeners have given equal rating to both $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$ for 42% of the test cases; and, for $\sim 32\%$ of the cases, the listeners have rated $\hat{s}_{x_{22}}(t)$ to be closer to reference than $\hat{s}_{x_{12}}(t)$. Also, the overall average rating for $\hat{s}_{x_{12}}(t)$ is very close to $\hat{s}_{x_{22}}(t)$ for most of the listeners. This is perhaps due to the inherent predisposition of performers of Hindustani music to elaborate individual notes, thereby minimizing the transitions between the 22 note positions.

Similar analysis on perception of Carnatic music shows that in 36% cases, ratings for $\hat{s}_{x_{12}}(t)$ and $\hat{s}_{x_{22}}(t)$ were the same. In $\sim 40\%$ cases, $\hat{s}_{x_{22}}(t)$ was found to be closer to reference than $\hat{s}_{x_{12}}(t)$.

In order to obtain measure of overall inter-listener agreement, we first categorize the ratings in each experiment into 3 categories - (i)$\hat{s}_{x_{22}}(t)$ closer to $\hat{s}_{ref}(t)$ (ii) $\hat{s}_{x_{12}}(t)$ closer to $\hat{s}_{ref}(t)$ (iii) equal ratings to both $\hat{s}_{x_{22}}(t)$ and $\hat{s}_{x_{12}}(t)$. For an i^{th} experiment, the agreement among the L number of listeners can be defined as [4]:

$$P_i = \frac{1}{L(L-1)} \sum_{j=1}^{3} n_{ij}^2 - L \tag{6}$$

where, n_{ij} is the number of raters who have assigned j^{th} category in i^{th} experiment.

In PT-2 perception test of Hindustani music, the average inter-listener agreement per experiment is found to be 0.37 while for Carnatic music, the average is 0.34.

From PT-1, we could infer that quantization at every pitch sample results in perceivable loss of *rāga* structure. The results of PT-2 shows that it is possible to quantize at critical points while retaining the *rāga* structure. There could be a few note omissions and distortions at micro-tonal levels which are not perceivable in one listening, implying *rāga* structure is well retained. This also implies that **quantizing critical pitch values keeps much of the *gamaka* structure** (which has been indefinable till now) **intact**. Expert musicians show sensitivity to 22-note positions; in some clips, musicians appreciate the approach to a note as interpolated by $\hat{s}_{x_{22}}(t)$ more than $\hat{s}_{ref}(t)$ [5]. We infer that both $K = 12 * 3$ and $K = 22 * 3$, depict close scores and retain *rāga* structure well.

[5] Sometimes, $\hat{s}_{x_{12}}(t)$ is also reported to interpolate transitions better than $\hat{s}_{ref}(t)$

3.5 Relation to Waveform Quantizers

The model corresponding to Equation (3) is a waveform quantizer. While an uniform quantizer assumes $y_n(t)$ to have uniform distribution, the model corresponding to Equation (1) and (3) is a non-uniform, parameterized, stochastic waveform quantizer. The stochastic SC-GMM incorporates shape of the pdf through its parameters to derive rendition-specific and/or *rāga*-specific quantization thresholds. While a well-designed optimum waveform quantizer with sufficient bit-depth can result in hi-fidelity audio, we have shown, from results of perception experiment PT-1, that non-uniform, parameterized pitch 'waveform' quantization unsettles the *rāga-bhava* even within a small 7 s melodic phrase. Increasing bit-depth without correlating to essential pitch-ratios (within an octave) will be of limited utility.

From model defined by Equation (2) and (4), we have seen from results of PT-2 that sub-sampling (at critical points) and then using a non-uniform, parameterized, stochastic quantizer results in melodic contours which can reconstruct *rāga-bhava* with less distortions. Increasing bit-depth (from $K = 12 * 3$ to $K = 22 * 3$) need not always result in lesser 'perceptual' distortions in melody signals which are inherently structured.

4. APPLICATIONS OF QUANTIZED PITCH CONTOUR

4.1 Note Transcription

A direct application of discretizing melody contours is in note transcription. While attempts have been made to capture regions corresponding to discrete notes, we now theorize that discrete notes can occur as points and/or regions in the melodic-temporal domain; elongated notes result in regions, while other-wise they can be essentially considered as points.

4.1.1 Experimental Setup

We have recorded a total of close to 50 phrases each in Hindustani and Carnatic music forms; the phrases are spread across 5 *rāgas* (as listed in Table 3) and is a modest database to quantify accuracy of note transcription. These phrases contain *rāga* specific *gamakas* such that their conventional transcription differs from their renditions. The Hindustani database is rendered with *Sārangi* instrument, while the Carnatic database contains vocally rendered phrases. Each phrase is associated with 2 note sequences - (i) note sequences as transcribed by musicians (referred as TA transcription) (ii) note sequences as musicians render it with the associated *gamakas*, but now explicitly notated (referred to as TB transcription). Figure 4 is a sample depicting the differences between TA and TB. The transcription notation used here consider only the note sequences and do not include duration information.

Pitch is extracted using Praat [2] every 8 ms. A SC-GMM model is built for each *rāga* by combining all phrases; note sequences are obtained as per Equation 4.

4.1.2 Results and Analysis of transcribed sequences

The performance of automatic note transcription task is measured using TA and TB transcriptions as ground truths. The automatically obtained note sequence is aligned with TA (or TB) using Needleman Wunsch global alignment algorithm [16] with gap penalty set to zero. Performance is reported in terms of recall accuracy and insertion rate. Table 3 summarizes the performance of automatic transcription using $k^*_{y_{12}}$ and $k^*_{x_{12}}$ note sequences in both Carnatic and Hindustani music.

Transcription using $k^*_{y_{12}}$ sequences always shows high recall results (as expected) and also results in high insertion rate; as every pitch sample is quantized, there is less likelihood of missing any note but more chances of false alarms.

With TA transcription as ground truth, recall rates using $k^*_{x_{12}}$ sequences is comparable to that using $k^*_{y_{12}}$ in Hindustani music; for Carnatic music, recall rate performance of $k^*_{x_{12}}$ sequence is seen to have decreased. Due to dynamic nature of Carnatic music, some notes in TA are not representative of the rendition. For example in *rāga Thodi*, though TA contains note Ga, it is rendered as $Ma-Ri$ (cf. Figure 1). Also, frequent notes or stable notes have dominant presence as Gaussian component (reflected as $\boldsymbol{\alpha}$); any critical pitch value in the vicinity of a dominant note can be assigned a higher probability in-spite of its distance to another adjacent but not-frequent note. This can cause incorrect pitch-to-note mapping.

Drastic reduction in transcription insertion rate can be attributed to $k^*_{x_{12}}$ sequences being estimated from sub-sampled version of $y_n(t)$. Thus, not all points are transcribed.

With TB transcription as ground truth, insertion rate is reduced for both $k^*_{y_{12}}$ and $k^*_{x_{12}}$ sequences. This is attributed to TB version of ground truth being a more elaborate explanation of a rendition. A sample depiction of the same can be seen in Figure 5. The number of false note assignment is reduced with transcription using $k^*_{x_{12}}$ as against $k^*_{y_{12}}$.

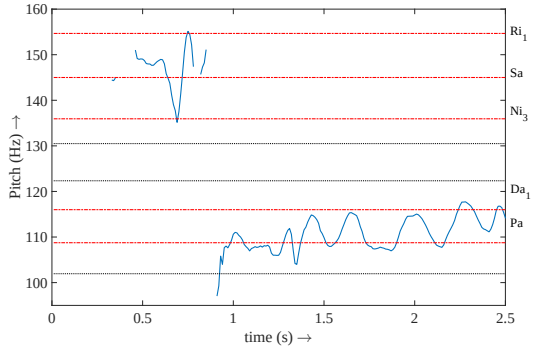

Figure 4. [Color Online] The blue contour corresponds to a phrase of *rāga Keervani* (Carnatic). Red lines indicate the pitch values of notes used in the *rāga*, while black lines denote pitch of notes that are not used. This phrase is transcribed as '$SaNiPaDa$' (TA). Considering the *gamakas* involved, it is rendered as '$SaNiRiSaPaDaPaDaPaDaPaDa$' (TB).

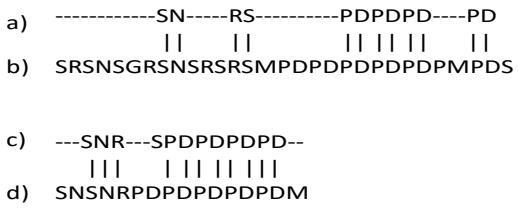

Figure 5. Melodic notes for pitch contour of Figure 4, corresponding to (a) Ground truth, TB (b) Transcription obtained using $k^*_{y_{12}}$ (c) Ground truth, TB (d) Transcription obtained using $k^*_{x_{12}}$

	(a) Hindustani							
	TA				TB			
	$k^*_{y_{12}}(t)$		$k^*_{x_{12}}(t)$		$k^*_{y_{12}}(t)$		$k^*_{x_{12}}(t)$	
Rāga	Rec	Ins	Rec	Ins	Rec	Ins	Rec	Ins
Bihag	1	3.65	1	1.55	0.93	1.39	0.88	0.39
Goud Sarang	1	4	1	2	0.88	1.97	0.83	0.88
Keervani	0.97	4.59	0.95	2.06	0.88	1.69	0.8	0.6
Madhuvanti	1	3.93	0.96	1.77	0.98	1.97	0.96	0.67
Marwa	1	7.18	0.97	3.36	0.96	2.4	0.88	0.90

	(b) Carnatic							
	TA				TB			
	$k^*_{y_{12}}(t)$		$k^*_{x_{12}}(\tau)$		$k^*_{y_{12}}(t)$		$k^*_{x_{12}}(\tau)$	
Rāga	Rec	Ins	Rec	Ins	Rec	Ins	Rec	Ins
Begada	1	8.58	0.85	1.91	0.98	2.93	0.80	0.32
Bhairavi	1	7.75	0.92	2.17	0.81	2.32	0.54	0.57
Hamsadvani	1	7.15	0.87	2	0.97	2.61	0.82	0.45
Hindola	1	10.8	1	3	0.9	3.03	0.86	0.46
Keervani	0.98	7.27	0.90	2.45	0.81	2.25	0.70	0.54
Thodi	1	10	0.88	2.55	0.92	3.04	0.88	0.36

Table 3. Performance of $k^*_{y_{12}}$ and $k^*_{x_{12}}$ sequences for automatic transcription in terms of average recall rate (Rec) and insertion rate (Ins) w.r.t. TA and TB ground truth transcription of phrases in (a) Hindustani (b) Carnatic music.

5. CONCLUSIONS

We have explored two different quantization techniques using stochastic models for mapping continuous melody contours to discrete pitch values; perception experiments show that *rāga-bhava* can be preserved by quantizing the pitch contour at critical points instead a waveform-quantization type of approach. The stochastic, parameterized SC-GMM assimilates information in pitch pdf to derive quantization thresholds. Applying results of perception experiments to automatic transcription task results in a detailed description of a melodic piece; such a detailed transcription can inherently aid in mapping *rāga* dynamics and *gamakas* into musical notation.

6. REFERENCES

[1] A. Bellur, V. Ishwar, and H. A Murthy. Motivic analysis and its relevance to raga identification in carnatic music. In *Proc. 2nd CompMusic Workshop*, pages 153–157, Istanbul, Turkey, 2012.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 6.0.14), 2016.

[3] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra. Essentia: An audio analysis library for music information retrieval. In *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, pages 493–498, 2013.

[4] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[5] K. K. Ganguli, S. Gulati, P. Rao, and X. Serra. Data-driven Exploration of Melodic Structures in Hindustani Music. *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, pages 605–611, 2016.

[6] K. K. Ganguli and P. Rao. Perceptual Anchor or attractor: How do Musicians perceive Raga Phrases. *Proc. Frontiers in Research in Speech and Music (FRSM)*, pages 174–178, 2016.

[7] S. Gulati, A. Bellur, J. Salamon, H. G. Ranjani, V. Ishwar, H. Murthy, and X. Serra. Automatic tonic identification in Indian art music: approaches and evaluation. *Journal of New Music Research*, 43(1):53–71, 2014.

[8] S. Gulati, J. Serrà, K. K. Ganguli, S. Şentürk, and X. Serra. Time-Delayed Melody Surfaces for rāga Recognition. In *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, pages 751–757, New York (USA), 2016.

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

[10] V. Ishwar, S. Dutta, A. Bellur, and H. A. Murthy. Motif spotting in an alapana in carnatic music. In *Proc. Int. Soc. on Music Info. Retr. Conf (ISMIR)*, pages 499–504, 2013.

[11] G. K. Koduri. *Towards a multimodal knowledge base for Indian art music: A case study with melodic intonation*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2016.

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

[13] A. Krishnaswamy. Pitch measurements versus perception of south indian classical music. In *Proc. of the Stockholm Music Acoustics Conference (SMAC)*, pages 627–630, 2003.

[14] S. McAdams. Psychological constraints on form-bearing dimensions in music. *Contemporary Music Review*, 4(1):181–198, 1989.

[15] Wim van der Meer and S. Rao. What you hear isn't what you see: The representation and cognition of fast movements in Hindustani music. In *Frontiers in Research in Speech and Music (FRSM)*, pages 1–8, Lucknow, India, 2006.

[16] S. B. Needleman and C. D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[17] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1993.

[18] H. G. Ranjani, S. Arthi, and T. V. Sreenivas. Carnatic music analysis: Shadja, Swara identification and Raga verification in Alapana using stochastic models. In *Proc. Workshop on Applicat. of Signal Process. to Audio and Acoust.*, pages 29–32, Oct 2011.

[19] G. Ruckert, A.A. Khan, and U.A. Khan. *The Classical Music of North India: The first years study*. Munshiram Manoharlal Publishers, 1998.

[20] P. Sambamoorthy. *South Indian Music*. Number v. 1-2 in South Indian Music. Indian Music Publishing House, 1963.

[21] S. Şentürk, G. K. Koduri, and X. Serra. A Score-Informed Computational Description of Svaras Using a Statistical Model. In *Proc. Conf. Sound and Music Computing (SMC)*, pages 427–433, Hamburg, Germany, 2016.