SINGING TRANSCRIPTION WITH SCHEMATIC NOTATION FOR RAGA MUSIC

Madhumitha S
DAP Lab, IIT Bombay

Govinda Madhava BS
DAP Lab, IIT Bombay

Preeti Rao
Professor, Dept of EE,
IIT Bombay

ABSTRACT

The use of a sparse representation for relatively complex music makes for a challenging automatic music transcription task in the context of the Indian art music traditions. We present a dataset that facilitates research on the association between the schematic notation and its vocal interpretation. For a set of available North Indian Khayal vocal performances in two ragas, we provide the canonical composition in schematic notation and also its realised notation as derived by an expert for each audio recording. We present our annotation pipeline, and investigate representations for both signal and notation in the course of developing a sequence model that is trained to predict the raga- and context-based notation from the extracted vocal pitch contour. Our limited and mixed results draw attention to the specific challenges we encountered.

1. INTRODUCTION

Notation in Indian art music traditions consists of note sequences, or *sargam* (solfa) syllables, written out aligned with the lyrics of the composition. The notes (*swar*) give an approximate guide to the melody with the precise rhythmic placement and pitch-continuous surface realization left to the performer based on their knowledge gained by oral transmission [1], [2]. This "schematic" notation is widely used in music pedagogy to convey a complex melodic phrase by singing its sequence of note names in place of the song's lyrics (as in the solmization we may see in a teaching video [1]). This is also very useful to instrument players learning to accompany singers. However, in order to use the Indian notation, one has to be familiar with the musical idiom. The implicit understanding of the intracacies of the raga and tala systems (i.e. the melodic and rhythmic framework) is what the expert musician brings to the act of writing out notation by ear for a given piece of vocal music or, conversely, interpreting the notation in order to sing a given pre-composed piece. This makes the association between notation and its vocal interpretation an attractive target for deep learning which, of course, ultimately points to the need for suitable annotated data.

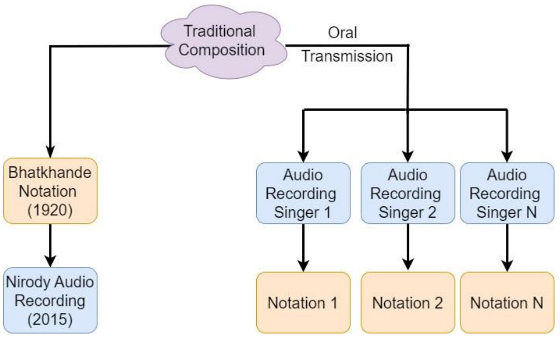


Figure 1. The bandish as the abstraction of a traditional composition with several different audio performances. Nirody's performances of Bhatkhande-notated compositions are available on CD [5].

Bandish are lyrical compositions in the North Indian, or Hindustani, music genre that serve as a reference for the associated raga in terms of the overall pitch movement and characteristic phrases [1]. The two verses of the bandish, termed *sthayi* and *antara*, typically sung early on in any concert, are defined by the lyrics and tune. The tune embodies the properties of a raga that a student needs to understand in order to sing, perform and eventually improvise in the raga. In the early 20th century, music scholar Pt. V.N. Bhatkhande launched on a mission to collect and notate traditional compositions from across the country. He devised his own ways of notating the melodies that so far lay within the framework of oral transmission and largely within hereditary musician families [3]. Now widely respected, his monumental work helped preserve the traditional compositions for posterity [4]. While several of the bandishes are regularly performed in the current day by artists across *gharanas* (schools), the surface realization in any performance could refer to a schematic version acquired by the artist through oral transmission and not necessarily the Bhatkhande book version. Figure 1 depicts this scenario, with the different schematic notations of the same (abstract) composition being of interest in our work.

In the context of MIR, we are solving a specific form of the note-level singing transcription task for a culture in which notation is schematic and skeletal, leaving out many important features of the music such as the inter-tonal nu-

[1] https://youtu.be/9qochgfPZhY?t=572

Attribution: F. Author, S. Author, and T. Author, "Singing Transcription with Schematic Notation for Raga Music", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.*, San Francisco, United States, 2024.

ances that are integral to the genre [6]. The main function of the notation is mnemonic and the flexibility this provides the performer with to incorporate specific styles or their own creative embellishments. All this, coupled with the practice of oral transmission, hints at potentially significant variations across the interpretations of a given composition across generations and artists. This is a characteristic shared with other oral traditions such as flamenco where experts have dedicated efforts to generating manual transcriptions from live performances [7]. In the case of raga music, the underlying schematic notation for any surface realization is linked to the corresponding raga's characteristics in terms of its melodic motifs and ornamentations [8]. This suggests the use of raga-dependent modeling for transcription and possibly new representations for both signal and notation that, further, do not depend critically on the availability of high-quality annotated data [9], [10].

A goal for our work is to eventually facilitate the automatic transcription to schematic notation of evolving versions of traditional compositions and of new compositions available only as recorded performances. Apart from impacting pedagogy, this will greatly enhance the scope of research in raga music in terms of exploiting the relatively vast audio resources for which notation has remained elusive so far [11], [12]. The chief contributions of this work are the dataset together with our carefully devised annotation process, and its validation via automatic transcription experiments.

The closest previous work to ours in terms of dataset design is the Carnatic Varnam Dataset of Compmusic [13]. This comprises 7 South Indian (Carnatic) music compositions (of the category called *varnam*, which is known for the strict adherence to the notation in performance), one in each of 7 distinct ragas. Apart from our Hindustani-focused work, we use multiple compositions in each raga based on the understanding that a more complete appreciation of a raga comes from familiarity with several different bandishes in the raga. We restrict ourselves to 2 ragas to provide the needed matched and unmatched test data contexts for experiments and potential tuning of deep learning models. A final, but crucial, difference is that while the Carnatic varnam audio recordings were created in the studio by getting professionals to perform the canonical form of the composition, our audio recordings (one or more per bandish) are obtained from available concerts (i.e. "in the wild"). This difference is particularly significant for Hindustani music given the overall lower reception of notation in this genre, the challenges of even manual notation due to the relatively loose coupling of the melody with the tala grid and the lack of standardization of the performance of traditional compositions when compared with Carnatic music [14].

In the next section, we describe our dataset in terms of its design and the annotation processes. This is followed by the presentation of methods for the automatic prediction of the notation for audio recordings of songs in the given raga. The results of our experiments are presented with a discussion of insights.

Full set of 12 tones	S r R g G m M P d D n N S
Raga Yaman	S R G M P D N
Raga Alhaiya-Bilawal	S R G m P D n N

Table 1. The pitch set of the middle octave, and that of each raga [15]. 'S' is the movable tonic. The full set of swar symbols across the 2 octave range is provided in the supplementary material site [2] .

2. DATASET

We choose ragas for our dataset based on the availability of a number of popular bandishes in the raga. This ensures the availability of audio recordings by professional musicians, accessible on internet. We avoid slow tempo (vilambit) bandishes due to the relatively high scope this presents for elaboration in the relatively long *matra* (tactus) intervals. Melodic aspects that distinguish ragas include the tonal material, the hierarchy of notes and their sequences leading to characteristic phrases [16]. The ragas we choose with their pitch sets are shown in Table 1.

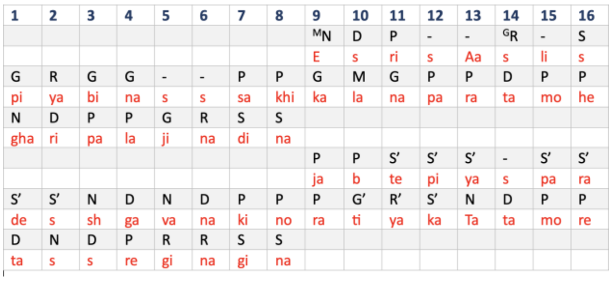
Central to our dataset is the bandish, a lyrical composition in the raga with two verses of 2-4 lines each. Each line of the text is set to a melody comprised of phrases of the raga. In concert, the artist typically introduces the bandish by singing it in its pre-composed form before beginning improvisation that entails structural modifications of the bandish lines, involving pitch and timing, apart from the smaller, expected variations due to expressiveness. With this in mind, we extract the data for this work from the initial part of the concert section corresponding to the bandish, with segment durations ranging from 1-3 min of the singing of the complete bandish at least once. We tag the cropped audio with laya, tala and singer/performance information.

Raga	Raga Yaman	Raga A-Bilawal
# Bandish	14	10
# Recordings	37	27
# Talas	2	2
# Artists	22	20
# Lines	545	384
# Notes (canonical)	8907	6611
Matra per min range	120-240	120-240
Duration (min)	65	41

Table 2. Overall summary of our dataset of recordings across ragas, bandishes and artists. More details provided at the supplementary link.

Table 2 summarises our dataset in terms of the audio recordings in each raga as performed by established artists. Each audio is associated with a traditional composition from Bhatkhande's book [4], where the canonical form in terms of lyrics and tune are provided in the rhythmic-cycle grid of the associated tala. While the artists in our dataset

[2] Supplementary material (anonymized): https://shorturl.at/lqstT



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
								^{M}N	D	P	-	-	^{G}R	-	S
								E	s	ri	s	Aa	s	li	s
G	R	G	G	-	-	P	P	G	M	G	P	P	D	P	P
pi	ya	bi	na	s	s	sa	khi	ka	la	na	pa	ra	ta	mo	he
N	D	P	P	G	R	S	S								
gha	ri	pa	la	ji	na	di	na								
								P	P	S’	S’	S’	-	S’	S’
								ja	b	te	pi	ya	s	pa	ra
S’	S’	N	D	N	D	P	P	P	G’	R’	S’	N	D	P	P
de	s	sh	ga	va	na	ki	no	ra	ti	ya	ka	Ta	ta	mo	re
D	N	D	P	R	R	S	S								
ta	s	s	re	gi	na	gi	na								

Figure 2. Notation and lyrics syllables for the two verses (4 lines) of bandish ’Eri Aali’ in raga Yaman placed in the 16-matra tala cycle grid with "1" specifying the sam or downbeat [4]

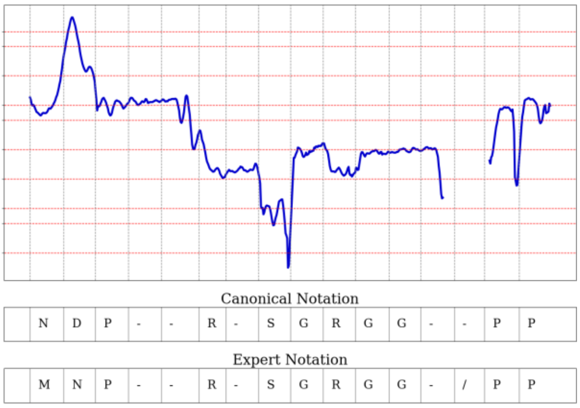


Figure 3. Top: F0 contour of the Line 1 of the bandish in Figure 2 performed by singer Nirody. Below is the aligned notation for canonical and expert (- denotes continuation, and / is silence) across 16 matra intervals. Horizontal lines mark the 100 cent pitch intervals nearest to the raga notes.

come from diverse backgrounds, we also have a set of performances by a senior musician, Indudhar Nirody, who performed all the bandishes as notated by Bhatkhande in a commercially available audio CD [5], [3]. Nirody’s songs are thus expected to be close reproductions of the published notation of Bhatkhande while the other artists could potentially deviate to some extent from the schematic notation due to referring to a possibly different source of the traditional bandish as suggested by their specific lineage.

Figure 2 shows a composition from Bhatkhande’s book [4]. Each ’line’ of the bandish here occupies 16 matras (beats) with each matra bearing a note label. We consider the main notes (with grace note in the superscript omitted since this would not appear in the sargam-syllable singing either) as specifying the schematic notation of the bandish. Next, for each of our audio recordings, we obtain the annotation of the performed audio by an expert musician with instructions to notate it as they would should they be asked to render the song in sargam (solfa) as they hear it. We term this the ’expert notation’. Trained musicians accomplish this familiar task by writing out short sequences of notes (melodic phrases), corresponding to roughly a few syllables of the lyrics at a time. This process of transcription, they do intuitively and quickly, in conformity with their knowledge of the raga’s tonal material and melodic contours. There is no attempt however to capture any precise note timing or ornamentation. Such an exercise would likely require repeated listening to possibly slowed-down versions to provide, what is known as, a *descriptive* transcription of the particular performance rather than just the underlying "blueprint" [6].

2.1 Audio processing

In computational studies of raga music, melodic shapes have been typically represented with continuous pitch (or F0) contours extracted from the singing voice [17]. Our audio recordings comprise the vocalist with drone and accompanied by tabla and, often, the harmonium. We use source separation [18] to obtain ’vocal’ and ’accompaniment’ tracks. The latter contains the tabla, drone and harmonium which is then processed for onset detection and beat tracking [19]. We estimate matra boundaries as the time instants aligned with those tabla strokes that are nearest to the globally estimated tactus pulses. Next, the vocal track is processed for pitch and voicing detection at 10 ms intervals using short-time autocorrelation analysis [20]. A critical analysis parameter is the pitch search range, limiting which can help optimise the analysis window and minimize octave errors. Other important parameters are the transition costs that trade off octave errors for accuracy in tracking the often large continuous pitch movements. Any gaps (up to 200 ms) in the pitch contour due to short unvoiced consonants or silences in the singing are interpolated. Due to the imperfections of source separation, we check for pitch detection errors via the sinusoidal resynthesis of the extracted F0 contour. A listening based comparison of the synthesized audio with the original audio mix quickly uncovers the automatic pitch detection errors based on which we choose to retain or reject the audio recording for our dataset.

The tonic, manually detected by a musician, was obtained for each performance and used to achieve the needed tonic-normalization of the extracted pitch contours. The detected tonic (i.e. *S* in the notation) is also used to define the expected roughly 2 octaves (i.e. from *’P* to *P’*).

2.2 Line-level segmentation and alignment

Each line of a bandish typically spans one tala cycle. As we observe from the example of Figure 2, the start of a bandish line does not necessarily align with the first matra of the cycle. The starting matra for each verse is noted with reference to the *sam* (first beat of the tala cycle). For each audio recording of the bandish, the lyrics are written out as continuous text for each of the tala cycles in the audio. Any skipped cycles are also indicated. This step does not require any music expertise. We use these lyric annotated audio segments as references for the associated canonical notation by automatically assigning a note label to each matra interval as obtained from the previous tabla-based markings. This step is followed with the similar process for the alignment of the expert notation (as obtained in Section 2). We observe that the notation of most of the audio recordings in our dataset comprise one note per matra in their assigned notation, although several have double notes and a few up to 4 notes. We have 51 instances of >2 notes per matra, out of approximately 8200 matra intervals. For our machine representation stage, we create a version with simplification of the matra-level 3- and 4-note groups to 2-note groups by retaining the first and final notes only. Further, note extension symbols as appearing in the ground-truth notation are replaced by the corresponding note symbol. Eventually, we have a set of audio segments, each corresponding to a line of a bandish, with its associated aligned notation - both the canonical form and the expert version, serving as two distinct types of ’ground truth’.

2.3 Data characteristics and organization

Given the availability of close to 900 performed bandish lines with parallel canonical and expert notations, our dataset affords a study of the structural differences between the traditional composition as recorded by Bhatkhande and the form realised by the singer.

On comparing the canonical and expert notations for audio segments across the dataset, we note that the highest similarity between the two occurs in the case of the performances by Nirody (as expected). The minor differences we see for this singer are exemplified by Figure 3, where the inflected note of Matra 1 is notated as *M* by the expert (instead of the canonical specified *N* note) given the visible drop in F0 in the interval. Matra 14 is marked with the blank symbol due to a singing pause. On the other hand, we also come across larger deviations between the canonical and expert notations for the same audio segment. These are attributed to differences in the gharana (school) and training of the performer, or to the intended improvisation for the bandish line. While we tried to select the first instance of bandish verses in the concert, the flexibility that marks Khayal singing and the pervasiveness of improvisation in all sections of the concert made it difficult to obtain sufficient data.

3. METHODS

Automatic music transcription, that was traditionally achieved with signal processing applied to quantize pitch and the timing of onsets, has been recently attempted mostly with deep learning models such as CNN and RNN [10]. We choose the BiLSTM, an RNN variant that incorporates both past and future context in its predictions. In this section, we discuss the representation of the input F0 time series as discrete tokens to the network, and the model architecture as depicted in Figure 4.

3.1 Representations

Our training data is in the form of labeled sequences. Each sequence corresponds to a line of the bandish spanning one rhythmic cycle of the audio. The choice of the line as a unit is based on the assumption that the necessary local context

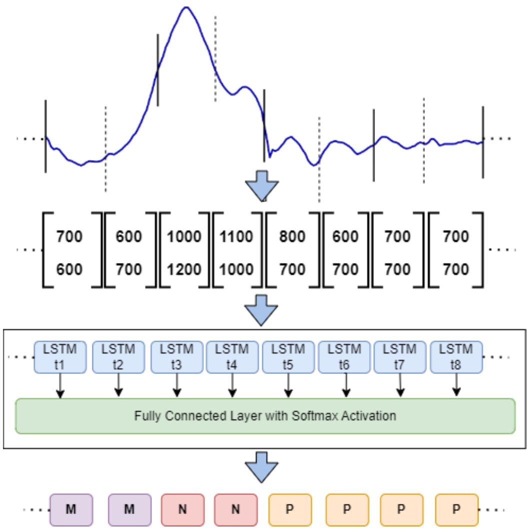


Figure 4. Top to bottom: F0 time-series with half-matra intervals marked via vertical lines. Each interval is represented by a 2-dim vector of medians (rounded to 100 cent multiples). The output of the BiLSTM and FC layers is a note sequence from the 26-alphabet symbol set. Note that the matra-level notation is expanded to fit the chosen tatum of half-matra in the presented 4-matra segment.

for assigning notation does not extend beyond the boundaries of a given line of the bandish. As mentioned earlier, our notations contain single- and double-speed notes. To account for note durations as small as one-half of a matra, we take the half-matra as the tatum and assign a note label to each tatum interval from the available matra-aligned sargam sequence based on the available notation for the line. The input to the BiLSTM network is then the F0 contour extracted from the cycle-duration audio signal segment. With the automatically detected matra boundaries, the F0 values within each half-matra segment are converted to input features associated with each tatum interval. We investigate different matra-level features in terms of aggregating the F0 samples within the interval in different ways.

Depending on the matra duration, we have a variable number of F0 samples (one every 10 ms) in the F0 contour that spans the matra, e.g. for a mpm of 120, we have 50 samples per matra, or 25 samples for our half-matra tatum interval. We represent these with a lower and fixed-dimensional vector. We first implement a piece-wise aggregate approximation (PAA) over the tatum-level F0 time-series, representing the tatum interval with one or more ’pieces’. The number of pieces is a design choice that trades off smoothness for the preservation of the contour shape in the interval. We test the following two configurations: 2 pieces per tatum interval to give us a 2 dimensional vector representation, and 1 piece to get a 1 dimensional representation. This is followed by assigning a discrete symbol to each resulting aggregate where the symbols are drawn from a suitable alphabet which comprises notes (swar) across the 2 octave range + silence. PAA is a time-series representation that has been used widely in data-mining tasks [21].

We do not use any separate note onset indicators given our assumption that the matra intervals are specified by the tabla-stroke onsets obtained from the tabla stroke track. This assumption does not hold for the matra intervals notated with double notes. Further, given the expressive singing style, the assumption may not strictly hold for the matra intervals either.

3.2 Network architecture

For the experiments, we train a model comprising a single BiLSTM layer having 32 LSTM blocks followed by a fully connected layer with softmax activation (refer figure 4). The output layer has 26 units corresponding to the vocabulary size (comprising symbols for 2 octaves range + silence). One crucial hyperparameter in our architecture is the output dimension of the LSTM which is set to 96 based on a preliminary search for the best hyperparameters for our task. We use tensorflow (version 2.11.0) for all our LSTM implementations. Training involves the optimization of the cross-entropy loss function computed at each epoch to iteratively update the network parameters. Input vector values are normalized to [-1,+1] which aid stable convergence during training, and all input-target pairs of sequences are padded to a length of 32. This becomes necessary when the tala cycle has fewer than 16 matras as for the few non-teentaal audios in our dataset. The batch size is fixed at 32. We utilize the Adam optimizer [22] with a learning rate of 0.001 and gradient clipping by value at 0.5 to address the issue of exploding gradients. During training, the validation loss is computed at each epoch and used to set the early stopping criterion. We report the best validation accuracy during training.

4. EXPERIMENTS AND RESULTS

We report the performance of our automatic notation system in terms of note classification accuracy with respect to each of the ground-truth notation schemes, i.e. Bhatkhande canonical form and the one obtained from experts listening to the audio recording. Each half-matra interval is considered correctly recognised if the predicted note (or silence symbol) matches exactly the aligned ground-truth notation for the interval. We disregard the onset detection component that is typically an important part of automatic music transcription tasks. This could add value to Khayal music transcription where vocal syllable onsets do not necessarily align with matra boundaries (as indicated by the tabla strokes), except possibly at the downbeat instants.

As a simple baseline for our automatic notation experiments based on sequence modeling, we consider an instantaneous rule-based method where the input representation comprising of the median of the F0 contour in the interval is assigned the raga note nearest to it. This, of course, does not take any of the context outside the tatum interval into account, although it does benefit from the prior knowledge of the raga notes.

4.1 Training and validation splits

Although we have data for two ragas, we presently report results for raga Yaman only. The training and validation sets comprise non-overlapping audio recordings, each comprising its full set of sung bandish lines. It is ensured that every bandish (i.e. the composition itself) is represented in the training set by one or more of its audio performances, given that this provides us with all the available unique note sequences of the canonical notation. The validation set is created out of the subset of audio recordings which have counterparts in the training set, i.e. the same bandish sung by a different artist. We note that while all the audio performances of a given bandish are assigned the same line-wise canonical notation, they are likely to have received distinct expert notations.

We create 6 different training-validation splits, each of which comprises validation audio recordings obtained as described above. The number of bandish audio lines in any single validation set is between 67 to 107, with the remaining (out of the 545 in Table 2) in training.

4.2 Results and discussion

We discuss our overall results. Table 3 displays the accuracies obtained with the simple baseline that assigns the nearest raga note to the median pitch computed on the tatum-interval. We see that the number of inaccurately predicted notes is close to 50%, with slightly superior performance with reference to the expert notation. We expect the latter to match the audio surface realization better especially in cases of significant structural differences in the artist's performance with respect to the canonical (i.e. Bhatkhande) version of the composition. On closer inspection, we find that a significant number of deviations from the expert notation are confined to within one raga note interval. The relatively high error rate for automatic annotation is observed to stem from timing alignment errors between the audio (represented by its F0 contour) and the ground-truth notation (whether canonical or by the expert). The surface realisation retains largely the same melodic shape but with clearly audible expressive timing differences leading to misalignment with the grid-marked notation. Further, while Bhatkhande notation is provided by the matra grid as indicated in Figure 2, the expert can capture only note clusters reliably but not the breakdown within or across matras.

Table 4 presents the performance of the BiLSTM network with its best hyperparameters. For the per half-matra (tatum interval) input representation, we observe that a single element to capture the pitch across the tatum interval works better than 2 elements representing separate sub-intervals. The resolution of pitch quantization across semi- and quarter-tone choices does not influence the outcomes. The BiLSTM network appears to be somewhat more successful in predicting the canonical ground-truth but worse than the rule-based baseline for the expert ground-truth. This is explained by the canonical notation training data which contains several repetitions of the bandish lines causing the network to memorise these target note sequences even in the face of audio-level diversity. The expert notation target sequences are far more diverse in themselves and probably would have helped learning by the network if the timing alignment was more reliable. We present some examples drawn from our dataset to illustrate these challenges in our supplementary material [2] .

Canonical	Expert
47.96	51.81

Table 3. Baseline system accuracy (%) across the validation set with the input representation of one F0 median per half-matra for each of the ground-truth annotations.

Input Representation	BiLSTM with CE Loss	
	Canonical	Expert
[2 dim, semitone]	53.77 (0.30)	48.95 (0.34)
[1 dim, semitone]	56.82 (0.07)	50.16 (0.45)
[1 dim, quartertone]	56.28 (0.45)	50.14 (0.28)

Table 4. BiLSTM-based system accuracy across the validation set, for the different input representations. We show the mean accuracy (%) and standard deviation across 3 training trials.

5. CONCLUSIONS AND FUTURE WORK

We presented the problem of deriving musical notation from a vocal performance as is conventionally used in Indian art music written forms and pedagogical settings. We sought a solution to the automatic notation task in the form of a sequence-to-sequence classifier for which powerful frameworks exist. In our chosen context of North Indian Khayal compositions as sung by prominent artists of the day, we encountered several challenges in setting up sufficient amounts of training data for the sequence classifiers. Recognizing that surface realizations of a traditional composition can deviate in nearly as many ways as there are artists and performances, we engaged in discussions with experts who helped us with annotation of the audio recordings via schematic notation, a skill they acquire in the course of their training. We found that the schematic notation deviated from the canonical version of the bandish particularly for prominent artists who also tended to use improvisation more extensively throughout the course of the performance. Structural changes such as altering the starting matra of the bandish lines or redistributing note durations over the course of the same word of the bandish lyrics are fairly common. Given the challenges this poses to obtaining aligned transcriptions, we have a case for exploring model training loss functions such as CTC (Connectionist Temporal Classification) that have recently been applied to tasks related to automatic music transcription that work with weakly labeled data [23], [24]. A second important aspect is to utilise more explicit signal cues to the singer's expressive timing such as spectral magnitude vectors for vocal syllable onsets in a system that predicts both note pitch and onsets.

Any successful outcomes of an automatic singing-to-schematic notation system holds great promise for musicological research given the available vast audio archives. This can also pave the way for increased understanding to benefit music generation which can do with some form of culturally meaningful discrete representation. Interesting ideas for pedagogical use also present themselves such as the combination of the generated notation with singing synthesis to achieve sargam singing demonstrations automatically.

6. REFERENCES

[1] S. Rao and P. Rao, "An overview of hindustani music in the context of computational musicology." *Journal of New Music Research*, vol. 43, no. 1, 2014.

[2] M. Schachter, "Structural levels in south indian music." *Music Theory Online*, vol. 31, no. B, pp. 74–89, 2015.

[3] "2000 classic compositions from bhatkhande on cd," https://scroll.in/article/726180/, accessed: 2024-04-10.

[4] V. Bhatkhande, *Kramik Pustaka Malika.* Sangeet Karyalaya Hathras, India, 2013.

[5] I. Nirody, "Samarpan: 2000 classical compositions performed on audio cd," Swarasankula Sangeetha Sabha, 2015.

[6] N. Magriel and L. du Perron, "The songs of khayal," Manohar Press, New Delhi, 2013.

[7] E. Gómez and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," in *Computer Music Journal*, June 2013, pp. 73–90.

[8] A. Morris, *Transmission and performance of Khayal compositions in the Gwalior gharana of Indian vocal music.* PhD thesis, Univ. Of London, S.O.A.S., 2004.

[9] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," in *IEEE Signal Processing Magazine*, January 2019, pp. 20–30.

[10] G. Peeters and G. Richard, "Deep learning for audio and music." *Multi-faceted Deep Learning: Models and Data, J. Benois-Pineau, A. Zemmari Edts, Springer.*

[11] A. Srinivasamurthy and P. Chordia, "Multiple viewpoint modeling of north indian classical vocal compositions," in *International Symposium on Computer Music Modeling and Retrieval (CMMR)*, Suzhou, China, 2012, pp. 383–396.

[12] J. C. Ross, A. Mishra, K. K. Ganguli, P. Bhattacharyya, and P. Rao, "Identifying raga similarity through embeddings learned from compositions' notation," in *International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 515–522.

[13] "Carnatic varnam dataset," https://compmusic.upf.edu/carnatic-varnam-dataset, accessed: 2024-04-04.

[14] B. Wade, "Music in india: The classical traditions," Manohar Press, 2001.

[15] S. Rao and W. van der Meer, *Music in Motion*. NCPA, https://autrimncpa.wordpress.com/about/, 2006.

[16] G. Koduri, S. Gulati, P. Rao, and X. Serra, "Raga recognition based on pitch distribution methods," *Journal of New Music Research*, vol. 41, no. 4, pp. 337–350, 2012.

[17] K. K. Ganguli and P. Rao, "A study of variability in raga motifs in performance contexts," *Journal of New Music Research*, vol. 50, no. 1, pp. 102–116, 2021.

[18] R. Hennequin, F. V. A. Khlif, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, 2020.

[19] M. Müller and F. Zalkow, "libfmp: A python package for fundamentals of music processing." *Journal of Open Source Software*, vol. 6, no. 63, 2021.

[20] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series." *Data Mining and knowledge discovery*, vol. 31, no. B, pp. 107–144, April 2007.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[23] J. Y. Wang and J. S. R. Jang, "Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, no. B, pp. 383–396, September 2023.

[24] Zalkow, Frank, and M. Müller, "Ctc-based learning of chroma features for score–audio music retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2957–2971, 2021.