STM JOURNALS

Trends in Electrical Engineering
ISSN: 2249-4774 (Online), ISSN: 2321-4260 (Print)
Volume 9, Issue 3
www.stmjournals.com

An Approach to Adding Knowledge Constraints to a Data-driven Generative Model for Carnatic Rhythm Sequence

Kaustuv Kanti Ganguli[1,*] ,Carlos Guedes[2]
[1]Post-Doctoral Associate , Division of Arts & Humanities, New York University. Abu Dhabi, UAE.
[2]Professor, Associate Arts Professor of Music, New York University. Abu Dhabi, UAE.

Abstract

Computational models for generative music have been a recent trend in AI based technology developments. However, an entirely data-driven strategy often falls short of capturing the naturally occurring rhythmic grouping. Guedes et al. [1, 2] had proposed dictionary based and stroke-grouping based approaches to generate novel sequences in the 8-beat cycle of Aditala. More recently an attempt of incorporating arithmetic partitioning, as conceived by performers, was made [3] to get rid of the drawback of the former model being failure to capture long-term structure and grammar of this particular idiom and being only successful in capturing local and short-term phrasing. One way of solving this issue would be to consider a rhythmic phrase as a gestalt i.e. to hypothesize three rationales: (i) a sequence of strokes, when played in a faster speed, behaves as an independent unit and not a mere compressed version of the reference; (ii) context influences the accent – the same phrase is played differently when as a part of a composition versus as a filler (ornamentation) during improvisation; (iii) phrases show a co-articulation effect – the gesture differs in anticipation of the forthcoming stroke/pattern. Initial experiments show that a time-compressed version of the reference phrase played in 4x speed sounds perceivably different from the same reference phrase played at 4x speed by the same musician. This indicates that there is a gestural difference in articulating the same phrase at different speeds. We extract timbral features to understand the differences, though there is a context-dependence that seems to have been captured in a supra-segmental way, motivating us to investigate prosodic features. This indicates that a syntactically correct sequence may not serve as a semantically plausible one to a musician's expectancy. As the qualitative evaluation of CAMel [1] involves expert listening, we believe, adding the proposed knowledge constraints would add to the naturalness, hence acceptability, of the generated sequences.

Keywords: *Carnatic rhythm, generative model, Gestalt principle, gestural modeling*

****Author for Correspondence*** E-mail: {kaustuvkanti,carlos.guedes}@nyu.edu

INTRODUCTION

The realm of music information retrieval (MIR) is broadly categorized into two complementary and seemingly mutually exclusive approaches, namely analysis and synthesis. A large part of synthesis in the recent era of artificial intelligence (AI) dedicates to building computational models for generative music. Literature in non-Eurogenetic music constitutes some of the earlier examples in the area of culture-specific music technologies. One early study [1] from our research group had the goal to develop expert systems that can reliably generate music in this style of Indian classical music, envisioning a contribution on two levels: (i) the creation of tools for lay audiences to interact with musical styles beyond the Western ones; and (ii) the automatic generation of unlimited amounts of data for training machine learning algorithms.

However, an entirely data-driven strategy often falls short of capturing the naturally occurring rhythmic grouping. Guedes et al. [1, 2] had proposed dictionary based and stroke-grouping based approaches to generate novel sequences in the 8-beat cycle of *Aditala*. Authors used an n-gram approach for modeling Carnatic percussion generation. The

size of n-grams was set to up to a five-gram to test how past information and size of accumulated memory could affect and change the generation process. The generation process used these data to generate new stroke events sequentially.

More recently an attempt of incorporating arithmetic partitioning, as conceived by performers, was made [3] to get rid of the drawback of the former model being failure to capture long-term structure and grammar of this particular idiom and being only successful in capturing local and short-term phrasing. The authors aimed to overcome these issues by introducing a new data-driven approach of modeling the *tala* cycle based on a set of arithmetic partitions that capture reliably the rhythmic structure of the *tala*, which led to developing an application that improves the generation of Carnatic rhythms and enhances the interaction of the user by adopting data visualization techniques during the generation.

Other literature on computational models on rhythm in Indian music has covered aspects of analysis [4, 5] which is essential to understand the top-down performance aspect, and thereafter can be imparted as knowledge constraints to the generative model. There are reported studies that tried to link theoretical concepts and performance practices [6–10] with a view to understanding how the concepts manifest in playing. There have been studies reported on how musicians perceive music [11-15] established through behavioral experiments. The aim of our current study is to analyze gestural differences of a same rhythmic pattern at different speeds through audio features.

The organization of the paper is as follows. In the next section, we give a brief background of the music concepts and terminology, alongside certain aspects of performance strategies that constitutes as the knowledge that we aim to impart to our data-driven generative model. In section 3, we list a set of hypotheses and experiments to (dis)prove them. Subsequently we discuss insights and conclude with potential future directions.

BACKGROUND AND DATA COLLECTION

The rhythmic framework of Carnatic music is based on the *tala*, which provides a structure for repetition, grouping and improvisation. The *tala* consists of a fixed time length cycle called *avartana* which is further divided into equidistant basic time units called *aksharas* (strokes). The rhythmic complexities of Carnatic rhythm are especially showcased during the solo or *taniavartanam*. In a concert, each percussion instrument (*Mridangam* and *Kanjira*) performs separately and then they trade off in shorter cycles with a precise question-answer like session, followed by a joint climactic ending. The concept of groupings is a fundamental building-block of Carnatic rhythm. There are certain rules that are followed by percussionists that allow this rhythmic generation to be more musically aesthetic, rather than just a series of groupings.

The data collection procedure for the current study was different from mainstream MIR task oriented dataset creation in the sense that the content was recorded with an intention of demonstrating the strategies that performers use which are hardly available in musicology texts. We believe that such a case study has the potential of bridging the gap between theory and practice, thereby help generative music to "sound more natural". A professional Carnatic percussionist Akshay Anantapadmanabhan, who also acts as a consultant to our research group MaSC (for details, visit *music and sound cultures: http://masc.hosting.nyu.edu/*), was consulted for ideation of the hypotheses. Further he recorded demonstrative pieces, both compositions and shorter excerpts, with *Mridangam* and *Kanjira*, as applicable. The recordings were made in a studio environment with three microphones in the presence of metronome. The *konokkol* (vocables of the *aksharas*) of the same were also recorded.

The corpus consisted of percussion solo compositions and groove patterns in aditala at

90 bpm tempo. The duration of each piece ranged from 40 seconds to 4 minutes. The pieces were manually (for details, visit *an automatic transcription is underway development, but we chose to use manually curated analysis for the current study*.) annotated by the author who is also a professional vocalist in Hindustani music with training over two decades. The groove pattern relevant to this study is the "*tha ri ki ta thom*" phrase which consists of 5 *aksharas*. The same phrase was recorded in three different speeds – the tempo was preserved at 90 bpm whereas the stroke density were shifted to 2x and 4x. The phrases at these three speeds were repeated 4 times for checking consistency. The performer had freedom to choose his own spontaneous gesture while playing at higher speeds.

HYPOTHESES AND EXPERIMENTS

Addressing the issue of gestural differences of playing the same phrase in different context/tempi, one way of solving would be to consider a rhythmic phrase as a gestalt i.e. to hypothesize three rationales: (i) a sequence of strokes, when played in a faster speed, behaves as an independent unit and not a mere compressed version of the reference; (ii) context influences the accent – the same phrase is played differently when as a part of a composition versus as a filler (ornamentation) during improvisation; (iii) phrases show a co-articulation effect – the gesture differs in anticipation of the forthcoming stroke/pattern. In this paper, we shall take up only the first one and computationally support the same.

We setup a preliminary experiment to test the first hypothesis by creating time-compressed versions of the reference phrase (played at 90 bpm) into 2x and 4x speeds on Ableton Live (for details, visit *https://www.ableton.com/en/live/*) software. The resampling is intuitive and hence we skip the discussion of the time-warping methodology. Figure 1(a) shows the reference "*tha ri ki ta thom*" phrase (1x speed) in terms of waveform and narrow-band spectrogram (FFT size of 128 samples at 44.1 kHz sampling, 50% hop, log-magnitude, linear frequency). Figure 1(b) shows the same plots for the originally played phrase in 2x speed and 1(c) shows the corresponding 2x time-compressed version. Figures 1(d) and 1(e) show the same for the 4x time-compression. On visual inspection, it is evident that a time-compressed version of the reference phrase played in 4x speed is perceivably different from the same reference phrase played at 4x speed. This indicates that there is a gestural difference in articulating the same phrase at different speeds.

We extracted timbral features to understand the differences between the articulation of the two versions at 4x rhythmic density: (i) recorded by the musician as shown in Figure 1(d) highlighted in red, and (ii) synthetically generated by time-compression as shown in Figure 1(e). The two features discussed in this work are chosen carefully, one each from time-domain and frequency-domain, namely RMSE envelope and spectral flux. The region of interest, as highlighted, leads to an inference that the performer did not intend to clearly resolve the 5 *aksharas* but rather treated the whole phrase as a gestalt. This motivated us to further investigate by extracting certain acoustic features as presented next.

Figure 2 shows a comparison of the timbral features with a view to capturing the difference in articulation in natural playing (top tier) versus the synthetically generated audio (bottom tier). The RMSE envelope on top tier (red) has broader peaks than that of bottom tier (blue), indicating less localized energy at the onset locations for the natural playing. This gives rise to a rolling sound for the spontaneous playing, indicating a different gesture altogether and not just time-compression of the reference phrase. The spectral flux on top tier (orange) has flatter (and less number of) peaks than that of bottom tier (indigo), indicating more drastic change across frames for the synthetically generated audio. This gives rise to a staccato nature of the onsets as opposed to a rolling sound.

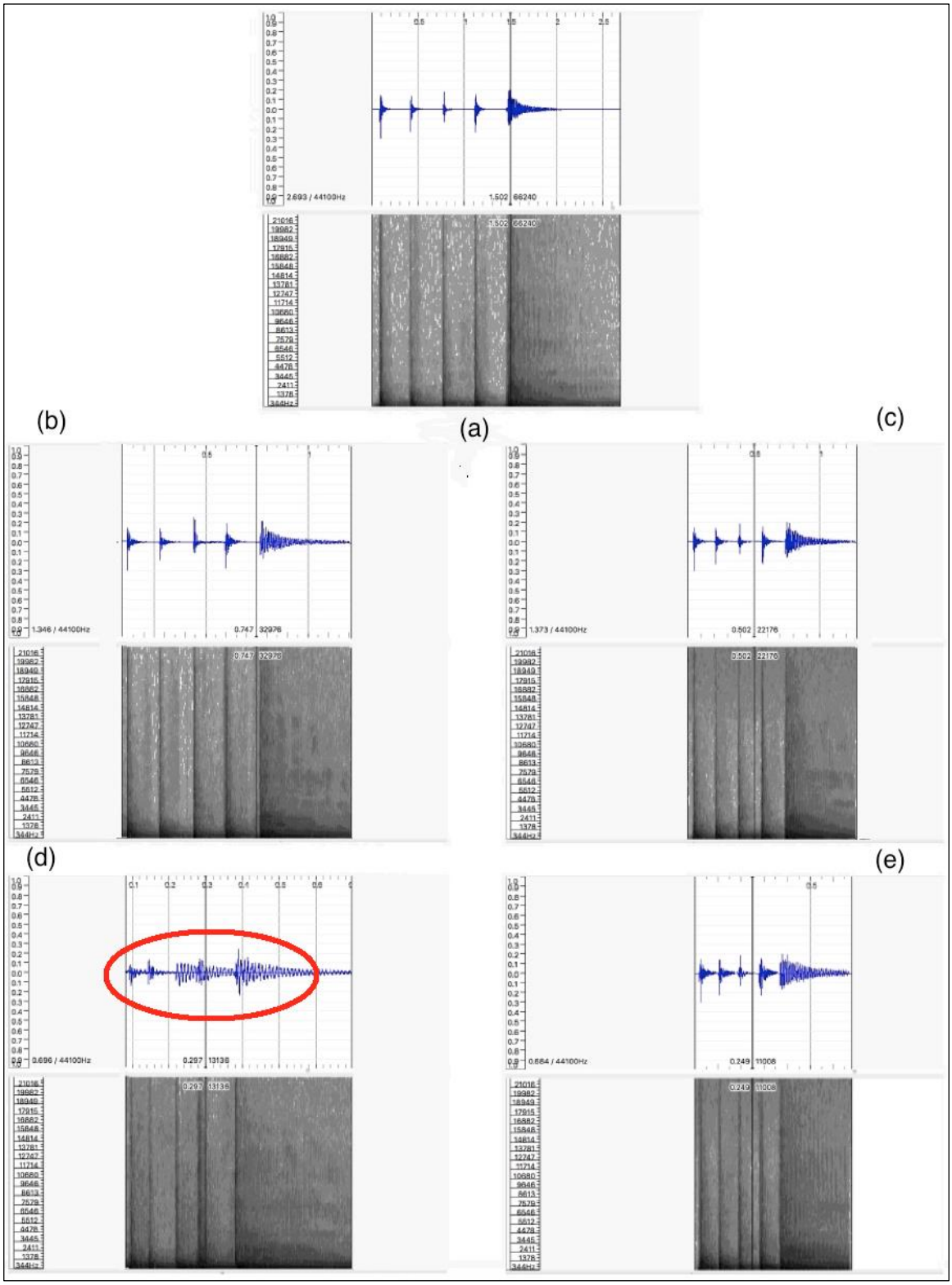

Fig. 1: *Waveform and spectrogram for different versions of the "tha ri ki ta thom" phrase under study. (a) reference phrase, (b) originally played phrase in 2x speed, (c) corresponding synthesized 2x time-compressed version, (d) originally played phrase in 4x speed, (e) corresponding synthesized 4x time-compressed version. The interesting observation lies in the difference between (d) and (e) – while (e) is intuitive, the waveform of (d) is highlighted to indicate the gestural difference. The durations are not to scale for better visual inspection.*

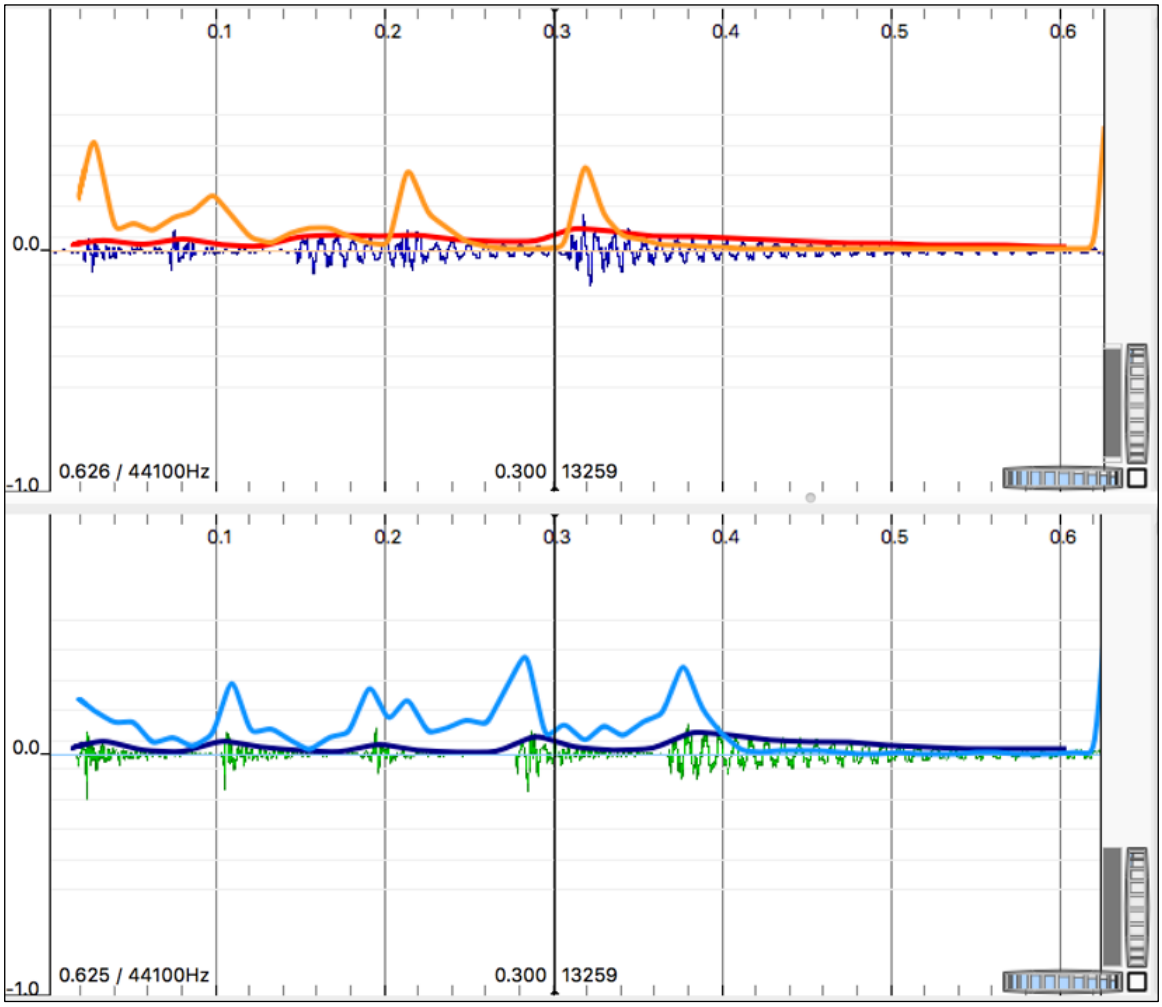

Fig. 2: *Comparison of the waveforms as shown in Figures 1(d) and (e). Timbral features RMSE envelope and spectral flux capture the difference in articulation of natural playing (top tier) versus the synthetically generated audio (bottom tier).*

DISCUSSION AND CONCLUDING REMARKS

Abiding by the basic thumb rules of engineering approaches, we quantified the segmental measures of RMSE and spectral flux for each ground truth segment *tha, ri, ki, ta, thom*. But it turns out that at 4x speed, the individual segments are not so important – so much that we cannot even clearly segment the 5 strokes from the natural playing. This is supported by the acoustic measurement as shown in Figure 2 (top tier) as we observe only 4 clear peaks in the RMSE and spectral flux. The third *akshara* "*ki*" seems to have the least prominent peak, though there is a slight difference between the two features. While addressing the possible way the performer's mental schema is manifested (and encoded) in the audio signal, we believe that there is a context-dependence which seems to have been captured in a supra-segmental way. This motivates us to further investigate prosodic features, and indicates that a syntactically correct sequence may not serve as a semantically plausible one to a musician's expectancy.

The context-based adaptation of hand gestures may be thought of as parallels with speech production concepts. Speakers modify vowel formants while changing the gesture from whisper through normal speech through shouting. This is to optimize power at certain spectral ranges by redistributing the energy. We find a similar notion that performers schematize to redistribute the weightage given

to individual strokes based on the given context.

We aim to add these supra-segmental features to the existing data-driven encoding of the generative model that can lead to solving the current issue of CAMel, i.e. handling of long-term sequences. However, it is to be noted that we have computationally modeled the data collected from only one performer. Though the hypotheses seem obvious, but these were constructed from discussion with only one model musician of the Carnatic repertoire. Given the diversity of the culture and existence of many schools of thought, the generalizability of the assumptions may be contested. A parallel ethnographic research through interviewing musicians could be a good means to (dis)prove the hypotheses. We can then propose new features or tweak the parameters of the computational analyses to support those.

Finally, an open ended question that still remains is what do we actually mean by a machine-generated music to "sound more natural" – it may be dynamic accent, more intuitive grouping, or even a better language model. As the qualitative evaluation of CAMel [1] involves expert listening, we believe, adding the proposed knowledge constraints would add to the naturalness, hence acceptability, of the generated sequences.

ACKNOWLEDGEMENTS

This research is part of project "Computationally engaged approaches to rhythm and musical heritage: Generation, analysis, and performance practice" funded through a grant from the Research Enhancement Fund at the New York University, Abu Dhabi, UAE.

REFERENCES

1. Guedes C, Trochidis K, Anantapadmanabhan ACAMel: Carnatic Percussion Music Generation Using N-Gram and Clustering Approaches. *Proceedings of 16th Rhythm Perception and Production Workshop (RPPW)*. Birmingham, UK. 2017.
2. Guedes C, Trochidis K, Anantapadmanabhan A. Modeling Carnatic Rhythm Generation: A Data-driven approach based on Rhythmic Analysis. *Proceedings of the 15th Sound & Music Computing Conference* Limassol, Cyprus. 2018.
3. Guedes C, Trochidis K, Anantapadmanabhan A. Challenges in computational modelling and generation of Carnatic percussion music. Challenges in Carnatic Music generation. *Proceedings of 5th International Conference of Analytical Approaches to World Music.* Thessaloniki, Greece. 2018.
4. Srinivasamurthy A, Holzapfel ,A, Serra X. In search of automatic rhythm analysis methods for Turkish and Indian art music. *Journal of New Music Research*. 2014; 43(1): 94–114.
5. Srinivasamurthy, A., Holzapfel, A., Ganguli, K. K., & Serra, X. (2017). Aspects of tempo and rhythmic elaboration in Hindustani music: A corpus study. Frontiers in Digital Humanities., 2017; 4: , 20.
6. Rao P, Ganguli KK. Linking prototypical, stock knowledge with the creative musicianship displayed in raga performance. *Proceedings of Frontiers of Research on Speech and Music (FRSM).* Rourkela, India. 2017.
7. Ganguli KK. How do we 'see' & 'say' a raga: A perspective canvas. *Samakalika Sangeetham*. 2013; 4(2): 112–119.
8. Ganguli, K. K., & Rao, P. (2018). On the distributional representation of ragas: experiments with allied raga pairs. Transactions of the International Society for Music Information n Retrieval (TISMIR). ,2018; 1(1): 79–-95.
9. Ganguli, K. K., & Rao, P. (2014). Tempo dependence of melodic shapes in Hindustani classical music. Proceedings of Frontiers of Research on Speech and Music (FRSM)., Mysore, India. 2014.
10. Ganguli KK, Rao P. Validating stock musicological knowledge via audio analyses of contemporary raga performance. *Proceedings of 20th Quinquennial Congress of the International Musicological Society*

(IMS): Digital Musicology Study Session. Tokyo, Japan. 2017.

11. Ganguli, K. K.,, & Rao, P. (2016). Exploring melodic similarity in Hindustani classical music through the synthetic manipulation of raga phrases. Proceedings of Cognitively-based Music Informatics (CogMIR). , New York, USA. 2016.
12. Ganguli, K. K., & Rao, P. (2015). Discrimination of melodic patterns in Indian classical music. Proceedings of National Conference on Communications (NCC), IEEE., Mumbai, India. 2015.
13. Ganguli KK, Rao P. Perceptual anchor or attractor: How do musicians perceive raga phrases? *Proceedings of Frontiers of Research on Speech and Music (FRSM)*. Baripada, India. 2016.
14. Ganguli KK, Rao P. Imitate or recall: How do musicians perform raga phrases? *Proceedings of Frontiers of Research on Speech and Music (FRSM)*, Rourkela, India. 2017.
15. Ganguli KK, Rao P. On the perception of raga motifs by trained musicians. *The Journal of the Acoustical Society of America*. 2019; 145(4): 2418–2434.

Cite this Article
Kaustuv Kanti Ganguli, Carlos Guedes. An Approach to Adding Knowledge Constraints to a Data-driven Generative Model for Carnatic Rhythm Sequence. *Trends in Electrical Engineering*. 2019; 9(3): 11–17p.