Questioning the Fundamental Problem-Definition of Mridangam Transcription

Kaustuv Kanti Ganguli[1†], Akshay Anantapadmanabhan[2] and Carlos Guedes[1]
[1] Music and Sound Cultures research group, New York University Abu Dhabi, UAE
[2] Freelance Musician, India
[†] Corresponding author: kaustuvkanti@nyu.edu

Introduction

There have been several attempts to analyze and characterize percussion instruments using computational methods, both in the context of Western (Sandvold et al., 2004; Tindale et al., 2004) and non-western percussion, specifically on automatic transcription of tabla (Chordia, 2005; Gillet & Richard, 2003) and mridangam strokes (Anantapadmanabhan et al., 2013; 2014). Although Anantapadmanabhan et al. (2013) provide greater insight into the mridangam strokes and their relation to the modes of the drumhead, the transcription approach is limited by its dependency on prior knowledge about the specific modes of the instrument. This puts a constraint for the method to be generalized to other instruments or different tonics (Anantapadmanabhan et al., 2014). Another concern is the unavailability of a unique mapping between the acoustic properties of a segmented stroke and its nomenclature in the vocabulary. It is often observed that the same stroke is uttered differently in the konakkol vocalization (the art of performing percussion syllables vocally in Indian art music), based on contextual variations or grammatical impositions. Most notably, even an expert musician is often unable to resolve such ambiguities on isolated presentation of a stroke. In this paper, we attempt to address this problem by proposing a combination of acoustic and semantic approaches for the contextual transcription of mridangam strokes.

We address the problem in an analysis-by-synthesis framework. First, a corpus of mridangam compositions is constructed and annotated. The annotations include both syntactic (i.e. an expert musician adhering to the lexicon without a reference to the acoustic properties of the audio) and listening-based (i.e. perceptual classification by a musician having no exposure to the mridangam repertoire). This facilitates modeling the task from both top-down and bottom-up approaches. In this work, we address the problem of mridangam stroke transcription at the intersection of these two approaches. The rest of the paper is structured in terms of description of the methodology, experimental results, and finally, discussion of the obtained insights.

Method

A corpus of mridangam compositions (both pre-composed traditional ones and spontaneous free-flowing) as well as konakkol was recorded by virtuoso percussionist Akshay Anantapadmanabhan (also a co-author). We employ traditional spectral methods (Bello et al. 2005; Peeters, 2010; Tindale et al., 2004) for transcription of the mridangam strokes detected by a spectral-flux based onset detection algorithm. To avoid the lexicon-based ambiguity, we initially performed a mid-level (manual) annotation based on 5 frequency bands: low, mid1, mid2, mid3, and hi. This was a reduction to five essential stroke types that can still represent the sequences according to Anantapadmanabhan. Figure 1 shows the ground truth onsets for the case-study excerpt. Comparing with the transcription in the form of konakkol, we see that there is a reduction from the set of unique konakkol syllables to the mridangam strokes. This leads to investigating the alias components in the transcription. This phenomenon occurs for two reasons: (i) to ensure fluency of reciting the konakkol at higher rhythmic densities, this is linked with physiological constraints of speech production; and (ii) to ensure the defined solkattu (rhythmic solfege for different subdivisions of the beat) to be recited in its integral form, this is linked to the language model.

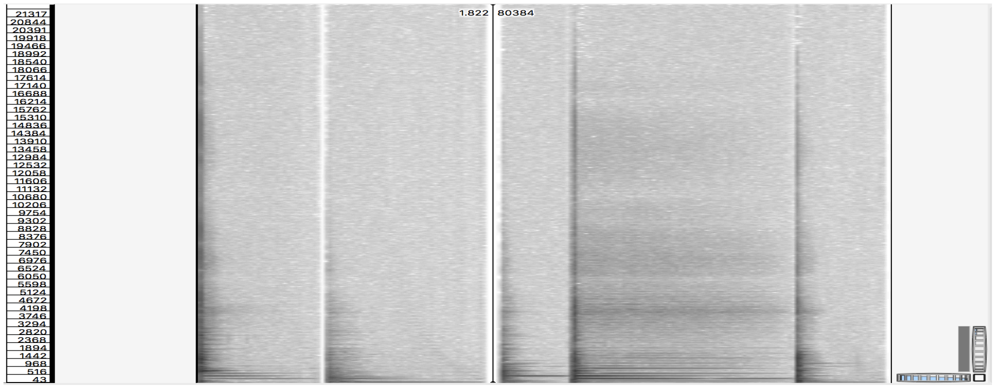

Figure 1: Spectrogram of typical strokes from the frequency-based classes [hi, low, mid1, mid2, mid3*] (left to right sequence) where the energy density is observed to vary along the frequency axis.*

Table 1 presents four typical phrases where the acoustically equivalent strokes are placed at different placeholders based on the grammar — mandated by ease of production for the konakkol. According to the mapping of konakkol-to-mridangam sounds, essentially Dhi ≡ Ka ≡ Ki — they are the same strokes in principle and are extensively used in the pedagogy to alert the pupils to realize the disambiguation process.

Table 1: Example phrases showing the acoustically equivalent strokes. The last two columns indicate the proportion of unique strokes to total stroke count without and with the acoustically equivalent strokes.

Index	***Phrases***	***Without equivalent***	***With equivalent***
1	*Ta Tha Cha Tha Ki Ta Tha Ka*	*5/8*	*4/8*
2	*Ki Ta Ki Ta Dhi Dhom Dhom Ka*	*5/8*	*3/8*
3	*Ki Ta . Ki Num . Ki Ta Ki Ta Dhom Dhom Ka*	*5/11*	*4/11*
4	*Tha . Dhi . Dhi . Thom Thom Ka . Dhi . Dhi Dhom Dhom Ka*	*5/11*	*4/11*

Thus we see that there is a further reduction in the set of (acoustically) unique[1] strokes after the introduction of acoustically equivalent strokes. This indicates that it is impossible to replicate the top-down (syntactic) transcription without the knowledge of this language model (mapping of the equivalents) into the algorithm.

Results

Given the 5 frequency-band based classes defined on the set of strokes are available as ground truth, the first approach we try is to use unsupervised clustering with the given number of clusters using a combination of frequency- and energy-based features. Figure 2 (left) shows the scatter plot of the detected onsets for a case-study excerpt with [Zero Crossing Rate (ZCR), Spectral Centroid] as the feature vector. The K-Means clustering with K=5 yields 5 clusters, denoted by different colors. All features are scaled between [-1,1]. For the second approach, we employ a simplistic supervised machine learning model K-Nearest Neighbor (KNN) classifier. We train the KNN with 30 strokes for each of the 5 classes (to ensure data balance) with a frequency-based (ZCR) and an energy-based (Spectral Centroid) feature. Figure 2 (right) shows the scatter plot of the ground truth onsets for the case-study excerpt. The predicted class labels are denoted by 5 different colors. Even though the homogeneity of the obtained clusters (left) are supported by a high cluster purity (0.97), there is a considerable overlap among the classified instances (right) that attracts further investigation on the mapping between the ground truth labels and the corresponding acoustic features.

[1] The definition of unique does not means identical in the signal domain, but similar within a small threshold. The standardization of similarity computation is an independent topic beyond the scope of this work. Without loss of generality, we can assume that the manner / place of articulation and hand gesture is identical.

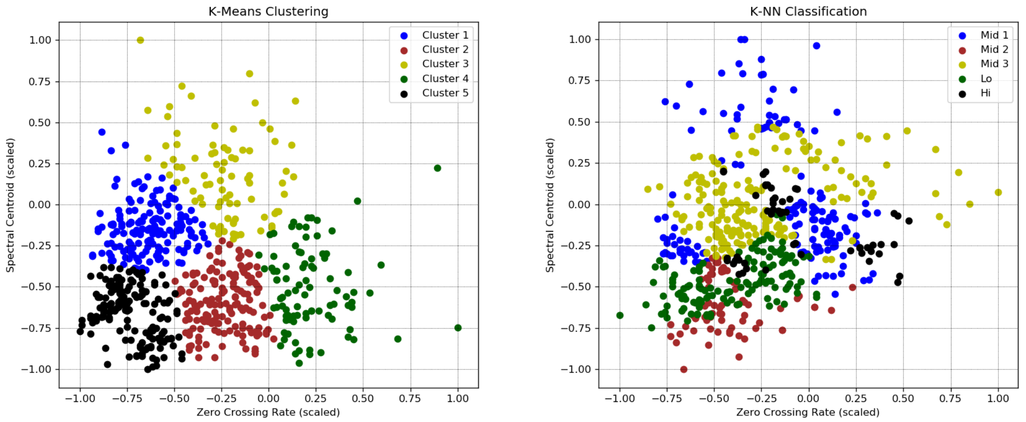

Figure 2: Scatter plots for K-Means clustering (left) and KNN classification (right) on the detected and ground truth onsets respectively, for a case-study excerpt.

The bottom-up computations show that a combination of machine learning approaches, i.e. unsupervised clustering (e.g. K-Means) and supervised classification (e.g. KNN), can lead to abstraction of frequency- and energy-based discrimination and labeling of mridangam strokes. However, the actual transcription labels (musicological) do not bear an intuitive mapping with the acoustic properties of the segmented mridangam stroke. This makes a two-pass system — where the knowledge constraints are to be incorporated in the post-processing stage — particularly relevant.

Discussion

We propose a 2-stage process for the assignment of the grammatically-accurate label for the transcribed mridangam stroke. In the first stage, we employed timbre-modeling techniques to assign a generic label for the stroke instance. In the second stage, the acoustically equivalent strokes are renamed according to the n-gram models learnt from the grammatical-annotation of the corpus as knowledge-constraints (see Ganguli & Guedes (2019)). This enables the mapping between konakkol-to-mridangam sounds adhering to the grammar prescriptions. Hence we contest the fundamental definition of the stroke-transcription problem. Even though the existing methods can reliably transcribe isolated strokes, the contextual transcription is still ill-defined and becomes particularly challenging. Moreover, there is no clear theory whether in the top-down transcription, konakkol syllables bear any correspondence with the frequency- or energy-characteristics of the corresponding strokes. Ganguli & Guedes (2019) showed that in the case of higher rhythmic density (tested on a phrase *Tha Ri Ki Ta Thom*, originally recorded at 90 bpm*)*, 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. Co-articulation effect mandates the performer to perceive the whole phrase as a gestalt (as opposed to a sequence of individual strokes) to modify the hand gestures which changes the acoustic properties of the realized stroke whereas the transcription label remains unaffected.

As future work, we propose word-vector (Mass et al., 2011) and graph-community (Blondel et al., 2008) based approaches to learn the context-dependence (e.g. language model) of the konakkol i.e. the symbolic subsequences. This will facilitate disambiguating the acoustically equivalent strokes from the detected (acoustically accurate) stokes from the first stage. Additionally, the network visualizations can in turn be fed back into the pedagogy as an interactive interface for engagement-learning.

Acknowledgments

This research is part of the 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.

References

Anantapadmanabhan, A., Bello, J., Krishnan, R., & Murthy, H. (2014). Tonic-independent stroke transcription of the mridangam. In Proceedings of Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. In IEEE Transactions on Speech and Audio Processing, 13(5), 1035-1047.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. In Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Chordia, P. (2005). Segmentation and Recognition of Tabla Strokes. In Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference, pp. 107-114, London, UK.

Ganguli, K., & Guedes, C. (2019). An approach to adding knowledge constraints to a data-driven generative model for Carnatic rhythm sequence. In Trends in Electrical Engineering, 9(3), 11-17.

Gillet, O., & Richard, G. (2003). Automatic labelling of tabla signals, In Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference, Maryland, USA.

Guedes, C., Trochidis, K., & Anantapadmanabhan, A. (2017). CAMeL: Carnatic percussion music generation using N-gram and clustering approaches. In Proceedings of the 16th Rhythm Production and Perception Workshop, Birmingham, UK.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 142-150. Association for Computational Linguistics.

Peeters, G. (2010). Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. IEEE Transactions on Audio, Speech, and Language Processing, 19(5), 1242-1252.

Sandvold, V., Gouyon, F., & Herrera, P. (2004). Percussion classification in polyphonic audio recordings using localized sound models. In Proceedings of the International Conference on Music Information Retrieval (ISMIR) conference, pp. 537-540, Barcelona, Spain.

Tindale, A. R., Kapur, A., Tzanetakis, G., & Fujinaga, I. (2004). Retrieval of percussion gestures using timbre classification techniques. In Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference, Barcelona, Spain.