To appear in: Proc. of Int. Soc. of Music Information Retrieval (ISMIR)
Suzhou (China), October 23-27, 2017

ACOUSTIC FEATURES FOR DETERMINING GOODNESS OF TABLA STROKES

Krish Narang **Preeti Rao**

Department of Electrical Engineering,
Indian Institute of Technology Bombay, Mumbai, India.
krishn@google.com, prao@ee.iitb.ac.in

ABSTRACT

The tabla is an essential component of the Hindustani classical music ensemble and therefore a popular choice with musical instrument learners. Early lessons typically target the mastering of individual strokes from the inventory of bols (spoken syllables corresponding to the distinct strokes) via training in the required articulatory gestures on the right and left drums. Exploiting the close links between the articulation, acoustics and perception of tabla strokes, this paper presents a study of the different timbral qualities that correspond to the correct articulation and to identified common misarticulations of the different bols. We present a dataset created out of correctly articulated and distinct categories of misarticulated strokes, all perceptually verified by an expert. We obtain a system that automatically labels a recording as a good or bad sound, and additionally identifies the precise nature of the misarticulation with a view to providing corrective feedback to the player. We find that acoustic features that are sensitive to the relatively small deviations from the good sound due to poorly articulated strokes are not necessarily the features that have proved successful in the recognition of strokes corresponding to distinct tabla bols as required for music transcription.

1. INTRODUCTION

Traditionally the art of playing the tabla (Indian hand drums) has been passed down by word of mouth, and documentation of the same is rare. Moreover, recent years have seen a decline in the popularity of Indian classical music, possibly due to the relatively limited accessibility options in todays digital age. While tuners are commonly utilized with melodic instruments, a digital tool that assesses the timbre of the produced sound can prove invaluable for learners and players of percussion instruments such as the tabla, in avoiding deep-seated deficiencies that arise from erroneous practice.

Based on the fact that there is an overall consensus among experts when it comes to the quality of sound (in terms of intonation, dynamics and tone quality) produced by an instrumentalist [2], Picas et al. [3] proposed an automatic system for measuring perceptual goodness in instrumental sounds, which was later developed into a community driven framework called good-sounds.org [4]. The website worked with a host of string and wind instruments, whose goodness broadly depended on similar acoustic attributes. We follow the motivation of good-sounds, extending it to a percussive instrument, the tabla, which has a sophisticated palette of basic sounds, each characterized by a distinct vocalized syllable known as a "bol". Further, in the interest of creating a system that provides meaningful feedback to a learner, we explicitly take into account the link between the manner of playing, or articulatory aspects, and the corresponding acoustic attributes.

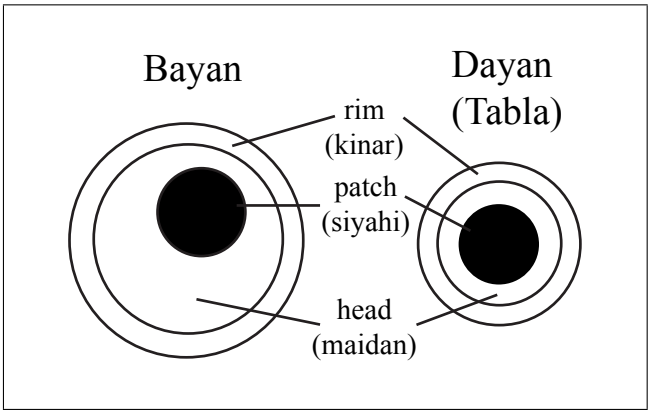

Figure 1. Regions of the left (bayan) and right (dayan) tabla surfaces, Patel and Iversen [1].

The tabla consists of two sealed membranophones with leather heads: the smaller, wooden-shell "dayan' (treble drum) is played with the right hand, and the larger, metal-shell "bayan' (bass drum) is played with the left. Each drum surface is divided into regions as shown in Figure 1. Unlike typical percussion instruments that are played with sticks or mallets hit at the fixed place on the drum surface, a tabla stroke is specified by the precise hand gesture to be employed (we term this the "manner of articulation", borrowing on terminology from speech production) and the particular region of the drum surface to be struck ("place of articulation"). Previous work has addressed the recognition of tabla bols for transcription via the distinct acous-

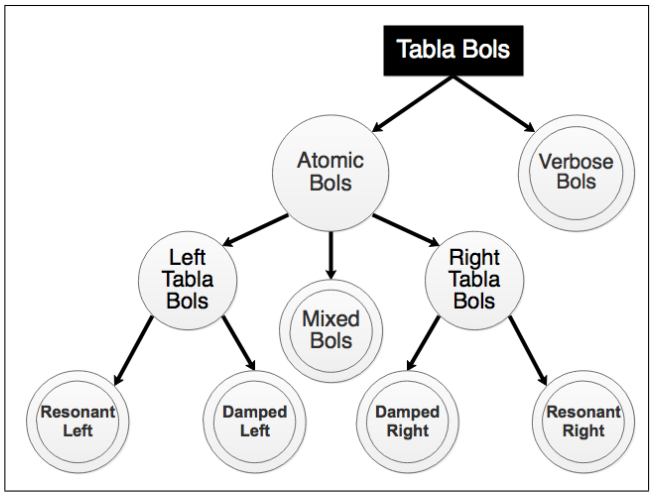

Figure 2. Articulation based classification of tabla bols.

tic characteristics associated with each of the strokes [5,6]. Temporal and spectral features commonly applied to musical instrument identification were used to achieve the classification of segmented strokes corresponding to different bols. Gillet and Richard [5] performed classification of individual bols by fitting Gaussian distributions to the energies in each of four different frequency bands. Chordia [6] used descriptors comprised of generic temporal as well as spectral features commonly used in the field of Music Information Retrieval for bol classification. More recently, Gupta et al. [15] used traditional spectral features, the mel-frequency cepstral coefficients, for the transcription of strokes in a ryhythm pattern extraction task on audio recordings. While the recognition of well-played strokes can benefit from the contrasting sounds corresponding to the different bols, the difference between a well-played and badly-played version of a bol is likely to be more nuanced and require developing bol-specific acoustic features. In fact, Herrera et al. [8] use spectral features for percussion classification based on a taxonomy of shape/material of the beaten object, specifically omitting instruments that drastically change timbre depending on how they are struck.

In this work, we consider the stroke classification problem where we wish to distinguish improperly articulated strokes from correct strokes by the analysis of the audio recording, and further provide feedback on the nature of the misarticulation. Based on a training dataset, that consists of strokes representing various kinds of playing errors typical of learners, as simulated by tabla teachers, we carry out a study of acoustic characteristics in relation to articulation aspects for each stroke. This is used to propose acoustic features that are sensitive to the articulation errors. Traditional features used in tabla bol recognition are used as baseline features and eventually we develop and evaluate a stroke classification system based on the combination of proposed and baseline features in a random forest classifier.

Type	Bol	Label	Position	Manner	Pressure
Resonant Left	Ge	Good	Maidan	Bounce	Variable
		Bad1	Siyahi	Bounce	Medium
		Bad2	Maidan	Press	Medium
		Bad3	Kinar	Bounce	Medium
Damped Left	Ke	Good	Siyahi	Press	Medium
		Bad1	Maidan	Press	Medium
		Bad2	Siyahi	Bounce	Light
Resonant Right	Ta/ Na	Good	Kinar	Press	Medium
		Bad1	Kinar(e)	Press	Heavy
		Bad2	Maidan	Press	Medium
		Bad3	Kinar	Press	Heavy
	Tun	Good	Siyahi	Bounce	None
		Bad1	Siyahi	Press	Light
		Bad2	Maidan	Bounce	None
	Tin	Good	Maidan	Bounce	Light
		Bad1	Siyahi	Bounce	Light
		Bad2	Maidan	Bounce	Heavy
Damped Right	Ti/ Ra	Good	Siyahi	Press	Medium
		Bad1	Siyahi	Bounce	Light
		Bad2	Siyahi(e)	Press	Medium
	Tak	Good	Maidan	Press	Medium
		Bad1	Maidan	Bounce	Light
		Bad2	Kinar(e)	Press	Medium
		Bad3	Siyahi(e)	Press	Medium

Table 1. Common articulations of bols in terms of position of articulation, manner of articulation, and hand pressure.

2. ARTICULATION BASED CLASSIFICATION

The tabla is a set of two drums, the left bass drum (bayan) and the right, higher pitched drum (dayan). Each tabla drum surface is composed of three major regions- siyahi, maidan, and kinar as depicted in Figure 1. Each tabla stroke (bol) is characterized by a very specific combination of the hand orientation with respect to the the position on drum surface, manner of striking, and pressure applied to the drum head, and has a very distinctive sound. Due to the heavy dependence of perceived quality of tabla bols on articulation accuracy of the player, it is instructive to understand the articulatory configurations of bols via the taxonomy visualized in Figure 2. Mixed bols are bols where both tablas are struck simultaneously (e.g. Dha, Dhin Dhit). Verbose bols (e.g. TiNaKeNa) consist of a sequence of strokes played in quick succession, whereas atomic bols are single stroke bols. A resonant bol is one where the skin of the drum is allowed to freely vibrate after it is struck, and a damped bol is one where the skin is muted in some way after it is struck.

Bols of each type (leaf nodes of Figure 2) can further be classified based on the place of articulation, manner of articulation and amount of hand pressure applied on the skin of the tabla. For example, for the bol tun, the index finger strikes the siyahi of the right tabla (dayan), with no damping (hand does not touch the tabla, finger is lifted after striking) (Patel and Iversen [1]). These are the three major attributes that distinguish bols within a type, and

are also what decide the perceptual goodness of a tabla stroke. For the same hand orientation, the drum can be struck sharply followed by immediately lifting the finger (we call this the 'bounce' manner of articulation) or it can be struck followed by leaving the finger or palm pressed against the drum head (we call this the 'pressed' manner of articulation). For the rest of the study we focus on atomic bols, which are sufficient for coverage of all beginner tabla rhythms, as listed on raganet, an educational magazine on Indian music [7]. For simplicity, mixed bols are not covered, since they are combinations of simultaneous left and right tabla strokes.

Two tabla teachers were consulted on the common mistakes made by beginners while playing a particular bol. Based on these, multiple classes were defined for each bol using the aforementioned three attributes governing goodness of a bol. One of these classes represents the well-played version of that bol, whereas the others represent the most common deviations that are perceptually distinct from the expected good sound. These are listed for all bols in Table 1, which explicitly shows the position, manner and hand pressure for different articulations of each bol, where "(e)" refers to the edge of the specified region. For example, a resonant right bol played on the maidan, while applying light hand pressure and lifting the finger after striking, constitutes a well-played Tin bol. However the same played while applying medium to heavy hand pressure is a badly-played Tin bol.

3. DATABASE AND ACOUSTIC CHARACTERISTICS

A dataset composed of 626 isolated strokes of 7 different bols was recorded (sampling rate of 44.1 kHz) by two experienced tabla players on a fixed tabla set that was tuned to D4 (294 Hz). The players were asked to play several instances of each stroke while also simulating typical errors that a new learner is likely to make in realizing a given stroke. Thus our dataset consists of recordings of each of bols realized in different ways as listed in Table 1, which also provides an articulation based description of the different realizations as executed by the tabla players. All the recordings were perceptually validated by one of the players who listened to each stroke and labeled it as "good" or "bad". In order to develop a system that provides specific feedback on the quality of a stroke, we required badly played instances of the bols as well. This made it impossible to use a publicly available dataset, as most archived recordings are from professional performances. Also, since our dataset is generated with reference to controlled variations in articulation as typical of a learner, it is likely to be more complete than the randomly sampled acoustic space of all possible productions.

A number of recordings was made per bol as seen in the Count column of Table 3, but with a roughly equal distribution of strokes across the classes corresponding to each bol in order to facilitate the construction of balanced training and test datasets for the classification task. The only exception to this is the bol Ge where a relatively large number of instances of the good stroke were produced since it is the only bol with pitch that can be modulated by changing the amount of pressure applied on the drum surface while striking. A number of such hand pressure based variations were recorded for the correct articulatory settings of the Ge stroke in order to get a reasonably representative dataset for the good quality bol Ge (124 out of the total of 187 Ge strokes in Table 3). This was important to ensure that the classifier we build is robust to pitch variations and other irrelevant changes caused by an increase or decrease in hand pressure.

Since each stroke presented in Table 1 is characterized by specific articulation (in terms of place of articulation, manner of articulation and amount of hand pressure), the acoustic variability is likely to cover more than one dimension. By studying the short-time magitude spectra (i.e. spectrograms) of the recorded bols, we were able to isolate the acoustic characteristics that distinguished the various classes of each bol. Time-domain waveforms and short-time magnitude spectra for two bols, Tin (a resonant right bol) and Ke (a damped left bol) are shown in Figure 3 and Figure 4 respectively. We observe that the rate of decay of the time-domain waveforms clearly discriminate the good from bad strokes. Further, the saliency as well as rate of decay of the individual harmonics (horizontal dark bands in the spectrograms) are seen to differ between the differently realised versions of each of the strokes. The resonant bol Tin is characterised by strong sustained harmonic components for good quality. In contrast, the damped bol Ke has a diffuse spectrum and rapidly decaying temporal envelope when realised correctly in Figure 4 top. A bounce in the hand gesture, on the other hand, degrades the stroke quality, contributing the prominent harmonics seen in the low frequency region of the bottom most bad stroke in Figure 4.

4. DEVISING FEATURES

From acoustic observations similar to those outlined in the previous section, across bols and goodness classes, we hypothesize that the strength, concentration and sustain of particular harmonics is critical to the quality of realization of a bol, especially for the resonant bols. Based on this, we propose and evaluate a harmonics based feature set which we call Feature set A. The features are designed to capture per-harmonic strength, concentration and decay rates. Harmonic based features are computed for each of the first 15 harmonics by extracting the corresponding spectral region by passing the signal through a narrow bandpass filter centered around that harmonic. These are important for resonant bols. The energy, spectral variance, and decay rate of each of the bandpass-filtered signals are computed. The decay rate is obtained as a parameter corresponding to an exponential envelope fitted to the signal. The energy and variance together constitute the strength of the harmonic, whereas decay rate represents how quickly that particular harmonic dies out. Spectral shaping features include variance, skewness, kurtosis and high frequency content. These features are extracted using Essentia [9], an open-

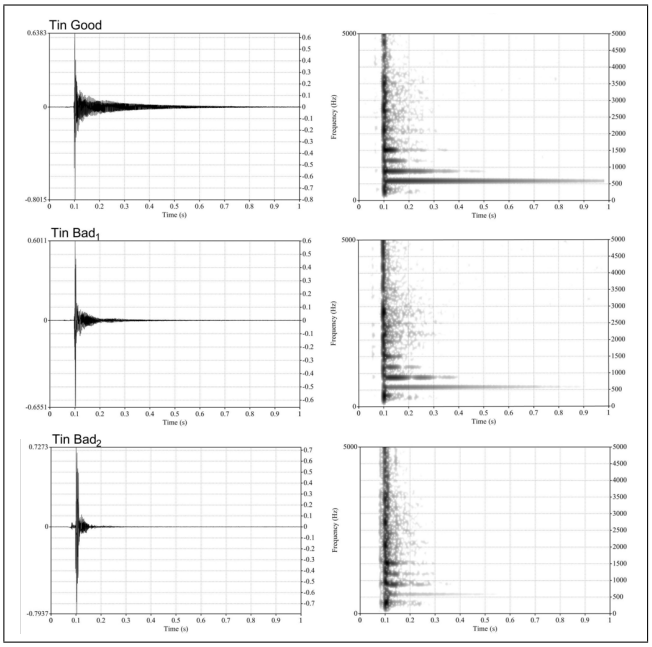

Figure 3. Waveform (left) and spectrogram (right) for good and selected bad recordings of Tin. Bad_1 is played in the wrong position, on the siyahi. Bad_2 is played with excess hand pressure.

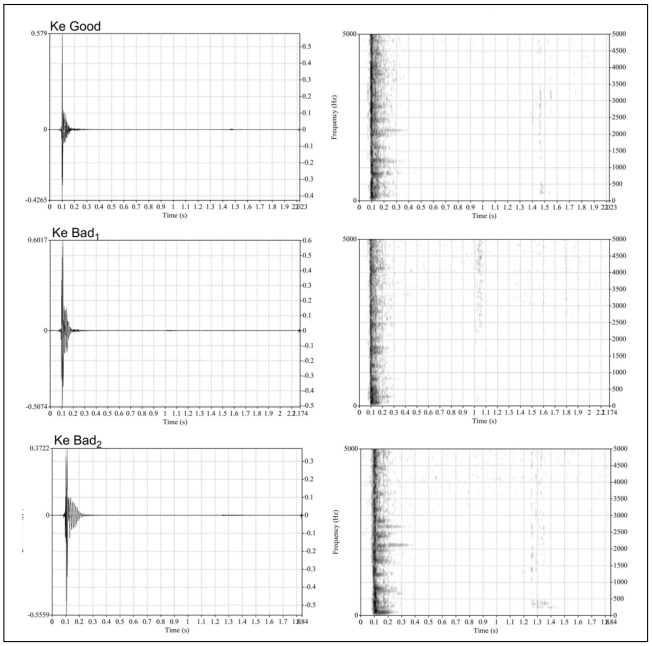

Figure 4. Waveform (left) and spectrogram (right) for good and selected bad recordings of Ke. Bad_1 is played in the wrong position, on the maidan. Bad_2 is played loosely- by bouncing the palm instead of pressing it.

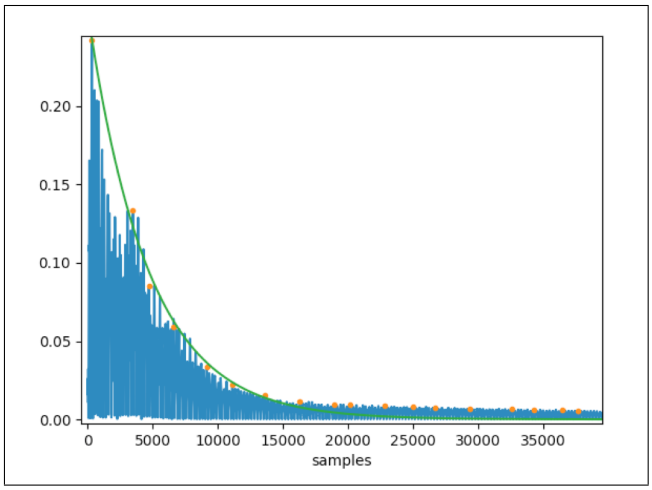

Figure 5. Exponential envelope fitted to rectified waveform for a Ge stroke. Dots mark the retained samples for curve fitting.

source library for audio analysis and audio-based music information retrieval. The temporal features include the energy and decay rate of the signal, and are useful for determining goodness of both damped and resonant bols. We also evaluate a baseline feature set (termed Feature Set B) which is essentially the same as the features employed by Chordia [6] in a tabla bol recognition task.

4.1 Harmonic Based Features

For each resonant bol that is correctly rendered, clear harmonics are visible in the spectrogram at multiples of a fundamental frequency. For resonant bols on the right tabla, the fundamental frequency is equal to the tonic of the tabla, except for Tun, for which the fundamental frequency is two semitones higher than the tonic [10]. However, these are not always precise, and a pitch detection algorithm should be used for determining the fundamental frequency of the recorded bol, e.g. the YinFFT algorithm [11]. For our dataset, the fundamental frequencies were manually estimated by viewing the spectrogram. For the tabla set used in our experiments, the tonic was determined to be 294 Hz, and fundamental frequency for Tun to be 330 Hz. For the left tabla stroke Ge the fundamental frequency was estimated to be 125 Hz.

For extracting harmonic based features, the signal is first passed through fifteen second-order IIR band pass filters with a bandwidths of 100 Hz and center frequencies at multiples of the fundamental frequency for that bol. Then an exponential envelope is fitted to the resulting time domain waveform. The waveform is full-wave rectified ($A'(t) = |A(t)|$), and only the maximum amplitude sample in every 50 millisecond interval is retained (as marked in Figure 5). The onset sample of the signal (assumed to be maximum amplitude sample over all time) is kept at $t = 0$. Next, SciPy's curve_fit function [12] is used to fit an exponential (ae^{-bt}) to the obtained samples, and both parameters a and b are considered as features. a represents the estimated maximum amplitude (referred in our

Bol	Selected Features
Ge	Energy(overall, 250, 500, 750, 1000, 1125, 1625), Decay(overall, 125, 250, 375, 625, 875), Impulse(125), Variance(125, 1500), MFCC(5, 6, 8, 10), Attack Time, Temporal Centroid, ZCR, Spectral Centroid
Ke	Energy(overall, 1764, 2352, 3528, 4116), Decay(overall, 294, 588, 2646, 3822), Impulse(294, 588, 882, 2352, 3234), MFCC(0, 1, 7, 12), Attack Time
Ta/Na	Energy(overall, 294, 1176, 1470), Decay(882), Impulse(2058), Variance(882, 1470, 2352), MFCC(1), Temporal Centroid
Tak	Energy(588, 882, 1176, 1470, 2646, 4116), Decay(294), Impulse(588), Variance(294), MFCC(1, 3), Attack Time, Temporal Centroid
Ti/Ra	Energy(588, 1764), Decay(588, 1176), Impulse(588), Variance(588), MFCC(11, 12), Attack Time, Temporal Centroid, ZCR
Tin	Energy(294, 2352, 3822), Decay(overall, 294, 588, 1470), Impulse(1764), Variance(294, 588), MFCC(2), Temporal Centroid
Tun	Energy(4950), Decay(330, 2310, 3960), Impulse(overall), Spectral Centroid, Temporal Centroid, ZCR

Table 2. Features selected from combination of set A and set B. The numbers in the bracket indicate the harmonic frequencies selected for energy/decay/impulse/variance and the indexes of selected coefficients (0-12) for MFCC.

feature set as 'impulse') of the signal and b represents the estimated decay rate (inversely proportional to the decay time). A similar curve fitting is done to the unfiltered time domain waveform. From the spectrum of the unfiltered signal, we calculate the energy and variance of the spectrum in bands centered around the first 15 harmonics with bandwidth equal to fundamental frequency. Finally the total energy of the signal is also taken as a feature. Finding energy and variance in a particular frequency range and band pass filtering were both done using routines from Essentia [9]. A total of 63 features were extracted in this way.

4.2 Baseline Feature Set

The baseline feature set consists of commonly used temporal and spectral features along with 13 MFCC's. These were used by Chordia [5] for tabla bol classification, and their relevance and effectiveness is also described in detail by Brent [13]. The temporal features are zero crossing rate, temporal centroid (the centroid of the time domain signal) and attack time. The attack time is calculated as time taken for the signal envelope to go from 20% to 90% of its maximum amplitude (default used in Essentia [9]). The spectral features are spectral centroid, skewness, and kurtosis. These are all obtained from the magnitude spectrum computed over the full duration of the recorded stroke. All of these features were computed using Essentia [9] routines.

Bol	Count	Classes	Set A	Set B	Combined Set
Ge	187	4	89.8	89.8	94.1
Ke	67	3	79.1	76.1	85.1
Ta/Na	86	4	89.5	86.1	91.9
Tak	101	4	80.2	82.2	86.1
Ti/Ra	79	3	77.2	96.2	91.1
Tin	48	3	89.6	93.8	97.9
Tun	29	3	81.0	91.4	93.1

Table 3. Percentage classification accuracies (one good class, multiple articulation based bad classes) Accuracies with Harmonic Features (Set A), Baseline Features (Set B), and selected features from a combination of Set A and Set B (Combined Set).

5. TRAINING AND EVALUATION OF BOL ARTICULATION CLASSIFIERS

Given our set of features, engineered as presented in the previous section, and the fact that our dataset is not very large, we employ a random forest classifier for the stroke classification task. A random forest classifier is an ensemble approach based on decision trees. We test for k-way classification accuracy in 10 fold cross validation mode with each of the different feature sets using the Weka [14] data mining software. Here k is the number of classes for a particular bol, consisting of one good class and multiple articulation based bad classes as shown in Table 3 where the number of strokes in the dataset for each bol is provided as well. For each instance the classifier predicts whether a bol is well-played (labeled good) or what misarticulations were made while playing the bol (labeled as the appropriate bad class). Apart from this, a subset of features is selected from the union of the two feature sets, using the CfsSubsetEval attribute selector with a GreedyStepwise search method from the Weka [14] data mining software. The greedy search picks each succeeding feature based on the classification improvement it brings to the existing set, using a threshold on achieved accuray as a stopping criterion. The set of selected features for each bol is shown in Table 2. Classification accuracies obtained with each of the 3 feature sets are presented in Table 3. We observe that the combination of features performs better than the baseline in nearly all cases. This indicates that the new harmonics based features bring in some useful information, complementary to the baseline features. In the case of the bol Ti/Ra, there is a decrease in classification accuracy with respect to the baseline. This is a damped bol and therefore harmonic features are not as important to it as spectral shaping features; however the issue of decreased accuracy after feature selection needs further investigation. Finally, Table 4 shows the results of two-way classification into good and bad strokes achieved by the combination features.

Bol	Feature Dimension	Accuracy
Ge	25	96.3
Ke	20	95.5
Ta/Na	11	96.5
Tak	13	94.1
Ti/Ra	10	92.4
Tin	12	97.9
Tun	8	93.1

Table 4. Percentage classification accuracies for two-way classification (good/bad stroke) based on features selected from the combined data set (as listed in Table 2).

6. CONCLUSION

Unlike many percussion instruments, the tabla is a musical instrument with a diverse inventory of basic sounds that demand extensive training and skill on the part of a player to elicit correctly. We proposed a taxonomy of strokes in terms of the main dimensions of articulation obtained through discussions with tabla teachers. This allowed us to construct a representative dataset of correct and common incorrectly articulated strokes by systematically modifying the articulatory dimensions. The results of this study show that nuanced changes in articulation are linked to perceptually significant changes in the acoustics of a tabla stroke. We presented acoustic features extracted from the isolated stroke segments to detect the articulation accuracy and therefore the perceptual goodness of a stroke from its audio. The best choice of features was observed to depend on the nature of the bol.

The present dataset was restricted to a single tabla set. For future work we would like to continue this research using a larger database from more sources, and to include coverage of mixed bols. The latter would further require measurements of relative timing between the atomic strokes that make up the mixed bol. This study can also easily be extended to evaluate sequences of bols (talas) for beginners- by a combination of rhythm scoring and evaluation of segmented bols of the sequence individually. The concept of expression and emotion in the playing of the tabla, which is vital to intermediate and expert players, is however a much more open ended question, and further research will hopefully lead to a characterization of that problem as well.

7. ACKNOWLEDGEMENT

The authors would like to thank Digant Patil and Abhisekh Sankaran for their help in recording and evaluating the Tabla strokes dataset. We are also indebted to Kaustuv Kanti Ganguli for lending his expertise in Indian Classical music, to Hitesh Tulsiani for his help with feature selection algorithms and classifiers, and to Shreya Arora for editing and rendering of graphs and images.

8. REFERENCES

[1] A. Patel and J. Iversen. "Acoustic and Perceptual Comparison of Speech and Drum Sounds in the North Indian Tabla Tradition: An Empirical Study of Sound Symbolism," *Journal of Research in Music Education, vol. 46*, pp. 522–534, 1998.

[2] J. Geringe and C. Madsen. "Musicians ratings of good versus bad vocal and string performances," *Proc. of the 15th international congress of phonetic sciences (ICPhS)*, pp. 925–928, 2003.

[3] O. Roman Picas, H. Parra Rodriguez, D. Dabiri, H. Tokuda, W. Hariya, K. Oishi, and X. Serra. "A real-time system for measuring sound goodness in instrumental sounds," *Audio Engineering Society Convention 138*, Audio Engineering Society, 2015.

[4] G. Bandiera, O. Roman Picas, H. Tokuda, W. Hariya, K. Oishi, and X. Serra. "good-sounds. org: a Framework To Explore Goodness in Instrumental Sounds," *Proc. 17th International Society for Music Information Retrieval Conference*, pp. 414–419, 2016.

[5] O. Gillet, and G. Richard. "Automatic labelling of tabla signals," *Johns Hopkins University*, 2003.

[6] P. Chordia. "Segmentation and Recognition of Tabla Strokes," *ISMIR*, pp. 107–114, 2005.

[7] A Batish. "Tabla Lesson 8 - Some Popular Tabla Thekas," *Batish Institute of Indian Music and Fine Arts*, `http://raganet.com/Issues/8/tabla8.html`, 2003.

[8] P. Herrera, A. Yeterian, and F. Gouyon. "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," *Music and Artificial Intelligence*, pp. 69–80, 2002.

[9] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra. "Essentia: An Audio Analysis Library for Music Information Retrieval," *ISMIR*, pp. 493–498, 2013.

[10] C. V. Raman. "The Indian musical drums," *Proc. of the Indian Academy of Sciences - Section A*, pp. 179–188, 1934.

[11] P. M. Brossier. "Automatic annotation of musical audio for interactive applications," *Diss. Queen Mary, University of London*, 2006.

[12] E. Jones, T. Oliphant, P. Peterson and others. "SciPy: Open Source Scientific Tools for Python," `http://www.scipy.org/`, 2001.

[13] W. Brent. "Physical and perceptual aspects of percussive timbre," *UC San Diego Electronic Theses and Dissertations*, 2010.

[14] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. "Data Mining: Practical machine learning tools and techniques," *Morgan Kaufmann*, 2016.

[15] S. Gupta, A. Srinivasamurthy, M. Kumar, H. A. Murthy, X. Serra. "Discovery of Syllabic Percussion Patterns in Tabla Solo Recordings," *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.