Four-way Classification of Tabla Strokes with Models Adapted from Automatic Drum Transcription

Rohit M. A., Amitrajit Bhattacharjee, Preeti Rao

Dept. of Electrical Engineering
I.I.T. Bombay, India
The Tabla

- Pitched, percussive hand-drums
  - Bayan - bass (F0 ∈ 80-100 Hz)
  - Dayan - treble (F0 ∈ 200-400 Hz)

- In performance
  - Solo – playing improvisation, compositions
  - Accompaniment – cyclic stroke pattern (*theka*)

[Image of Tabla instruments: Bayan (Bass) and Dayan (Treble)]

Tabla Strokes and Categories

- Tabla strokes
  - 10-15 in number, identified by ‘*bols*’ (syllables like Na, Tin, Ghe, Dha, etc.)
  - May involve single drum or both simultaneously
  - Resonant (R) – sustained, harmonic
  - Damped (D) – transient, percussive

- 4 stroke categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Bass drum</th>
<th>Treble drum</th>
<th>Bols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resonant Bass (RB)</td>
<td>R</td>
<td>D / Nil</td>
<td>Ghe, Dhe, Dhi</td>
</tr>
<tr>
<td>Resonant Treble (RT)</td>
<td>D / Nil</td>
<td>R</td>
<td>Na, Tin, Tun</td>
</tr>
<tr>
<td>Resonant Both (B)</td>
<td>R</td>
<td>R</td>
<td>Dha, Dhin</td>
</tr>
<tr>
<td>Damped (D)</td>
<td>D / Nil</td>
<td>D / Nil</td>
<td>Ti-Ta, Te-Re, Ke</td>
</tr>
</tbody>
</table>
Relevance of Stroke Categories

• Musicologically motivated
  • Mark salient positions in theka
  • Are tied to expressive tabla playing elements (loudness dynamics, pitch modulation) \(^2,^3\)

• Aiding computational musicology
  • Analysis of played strokes requires expensive manual labelling
  • Automatic stroke classification can enable corpus-level analysis

\(^3\) A. Srinivasamurthy et al. “Aspects of tempo and rhythmic elaboration in hindustani music: A corpus study,” Frontiers in Digital Humanities 2017
Existing Methods do not Generalize

• Previous work – tabla bol classification

• Poor instrument-independent classification accuracies
  • Highly variable test set accuracies on unseen tabla (15 - 95% ⁴)
  • Lack of sufficient data with diversity in playing & instrument characteristics

---

⁴ P. Chordia, “Segmentation and recognition of tabla strokes,” ISMIR 2005
Objectives

• 4-way tabla stroke classification system
  • Robust to instrument and playing-style changes

• Target classes
  • Damped (D)
  • Resonant Treble (RT)
  • Resonant Bass (RB)
  • Resonant Both (B)

• Target test scenario
  •Tabla accompaniment to Hindustani classical vocals
Approach

1. Build larger, diverse dataset

2. Design effective classification models
   - Exploit pre-trained models from western drums transcription

3. Explore novel data augmentation methods
Dataset

- New labelled dataset diverse in terms of
  - Instruments, players, tabla tuning, playing tempo

- Stroke category distribution is not uniform
  - Most strokes of D, least of RB

<table>
<thead>
<tr>
<th></th>
<th># tablas</th>
<th>Duration</th>
<th># strokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>5,6</td>
<td>10</td>
<td>76 min.</td>
</tr>
<tr>
<td>Testing</td>
<td>6</td>
<td>3</td>
<td>20 min.</td>
</tr>
</tbody>
</table>

Data Collection and Annotation

• Testing set
  • Realistic tabla accompaniment to vocals (theka)
  • Recorded in isolation

• Training set
  • Tabla solo audios (kaida, tukda, tihai, etc.)
  • Split into 3 cross-validation folds (no tabla overlap)
Methods

1. Classification – CNN models inspired from western automatic drums transcription (ADT)
   A. 3-way drums CNN with transfer learning
   B. Bank of retrained 1-way CNNs

2. Data augmentation
Overview of CNN methods

• Input
  • Mel-spectrogram excerpt of ‘C’ channels x ‘F’ bins x ‘T’ frames
  • Channels are spectrograms computed at different resolutions

• Target
  • Binary value indicating presence of onset at frame T/2

• Output
  • Onset probability in [0,1]
  • Multi-label: single model, multiple outputs
  • 1-way: different models, one-versus-all binary output

Figure inspiration: R. Vogl et al, “Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks,” ISMIR 2017
A. 3-way CNN with Transfer Learning

• Pre-trained CNN drums transcription (ADT) models from madmom\(^7\) fine-tuned on tabla data
  • Model originally trained on MIREX drums transcription dataset (about 3x our tabla dataset)

• Motivated by correspondence between drum types and tabla stroke categories

\(^7\) R. Vogl and P. Knees, “Mirex submission for drum transcription 2018”. ISMIR 2018
A. 3-way CNN with Transfer Learning

- Evaluate against re-training same model architecture on tabla data
B. Bank of 1-way CNNs

- Bank of four 1-way CNNs
  - Allows optimizing model separately for each category

- We start with a baseline architecture from ADT 8
  - Hyperparameters tuned for each category to account for data imbalance
  - Include variations to input representation and model capacity

- Trained from scratch for each category
  - No pre-trained ADT model available
  - Drums dataset used for 3-way CNN also not public

---

8 C. Jacques and A. Röbel. “Automatic drum transcription with convolutional neural network”. DAFx 2018
Methods

1. Classification – CNN models inspired from western automatic drums transcription (ADT)
   A. 3-way drums CNN with transfer learning
   B. Bank of retrained 1-way CNNs

2. Data augmentation
Data Augmentation - Overview

• Diversity expected in tabla dataset
  • **Instrument characteristics** – tuning, timbre, resonance & decay levels
  • **Playing style** – tempo, expressive dynamics
  • **Recording conditions** – spectral levels, balance, decay level

• We explore individual methods and combinations
  • Each method applied to time-domain training set audio, generates 4 variations
  • Evaluated using 1-way CNN models
Augmentation Methods From Literature

- Pitch-shifting (PS) & time-scaling (TS)
  - Capture tuning and tempo variations

- Attack remixing (AR)
  - Modify relative levels of signal attack and decay
  - Used previously in ADT

---

Tabla-specific Methods – Spectral Filtering

• Spectral filtering commonly used in MIR\(^{10}\)
  • Filter applied over randomly chosen spectral bands

• Given specific bands of activity in tabla, we filter bass & treble regions
  • Capture recording conditions, resonance characteristics

• We also identify and modify features that vary across instruments & not stroke categories
  • Perturbing attributes irrelevant to discrimination task shown to be effective for augmentation\(^{11}\)
  • For our task, these can be instrument-specific low-level acoustic features

Finding instrument-dependent characteristics

- **Tabla identification task** using a random forest classifier
  - About 50 features used, represent various spectral and temporal characteristics
  - Training set has audio from 10 unique tabla-sets
  - Samples from each stroke category are used separately to fit RF models

- Resulting feature ranking highlights MFCC-1 as most important

---

Tabla-specific Methods – Stroke Remixing

• Expressive playing involves modifying relative stroke intensities
• Compound strokes differ in contribution of each drum

• Decompose tabla audio into components of each stroke type
  • NMF with pre-initialized and fixed bases
• Remix at different levels to simulate playing dynamics

---

Results

1. Transfer learning with 3-way drums CNN models
2. Re-trained 1-way CNN models (bank of 4)
3. Data augmentation
4. Overall comparison
1. Adapted 3-way drums CNN

- Disjoint tuning of both conv and dense layers better than other fine-tuning approaches
- Higher mean f-score than re-trained model of same architecture

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroke category</td>
<td>D</td>
<td>RT</td>
<td>RB</td>
<td>B</td>
<td>D</td>
<td>RT</td>
<td>RB</td>
<td>B</td>
</tr>
<tr>
<td>Pre-trained (PT)</td>
<td>36.8</td>
<td>15.1</td>
<td>9.8</td>
<td>7.3</td>
<td>36.8</td>
<td>15.1</td>
<td>9.8</td>
<td>7.3</td>
</tr>
<tr>
<td>Re-trained</td>
<td>81.0</td>
<td>53.7</td>
<td>15.7</td>
<td>63.0</td>
<td>81.0</td>
<td>53.7</td>
<td>15.7</td>
<td>63.0</td>
</tr>
<tr>
<td>FT dense random init.</td>
<td>74.4</td>
<td>55.9</td>
<td>33.6</td>
<td>63.4</td>
<td>74.4</td>
<td>55.9</td>
<td>33.6</td>
<td>63.4</td>
</tr>
<tr>
<td>FT dense PT init.</td>
<td>71.7</td>
<td>54.8</td>
<td>29.4</td>
<td>60.9</td>
<td>71.7</td>
<td>54.8</td>
<td>29.4</td>
<td>60.9</td>
</tr>
<tr>
<td>Uniform FT all</td>
<td>76.3</td>
<td>59.7</td>
<td>29.5</td>
<td>65.3</td>
<td>76.3</td>
<td>59.7</td>
<td>29.5</td>
<td>65.3</td>
</tr>
<tr>
<td>Differential FT all</td>
<td>72.5</td>
<td>58.7</td>
<td>30.0</td>
<td>63.5</td>
<td>72.5</td>
<td>58.7</td>
<td>30.0</td>
<td>63.5</td>
</tr>
<tr>
<td>Disjoint FT all: dense rand. init.</td>
<td>77.2</td>
<td>57.4</td>
<td>33.0</td>
<td>65.9</td>
<td>77.2</td>
<td>57.4</td>
<td>33.0</td>
<td>65.9</td>
</tr>
<tr>
<td>Disjoint FT all: dense PT init.</td>
<td>74.8</td>
<td>66.4</td>
<td>34.7</td>
<td>66.5</td>
<td>74.8</td>
<td>66.4</td>
<td>34.7</td>
<td>66.5</td>
</tr>
</tbody>
</table>
2. Re-trained 1-way CNNs

- Separate hyperparameter tuning results in superior class-specific architectures
  - D: more dense layer units
  - RT: more conv layer filters
  - RB and B (data scarce): Baseline model

<table>
<thead>
<tr>
<th>Cross-validation F-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroke category</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>↑context</td>
</tr>
<tr>
<td>Mid-channel</td>
</tr>
<tr>
<td>↑conv filters</td>
</tr>
<tr>
<td>↑dense units</td>
</tr>
<tr>
<td>↑conv filters+↑dense units</td>
</tr>
<tr>
<td>2x conv layers</td>
</tr>
</tbody>
</table>
3. Data augmentation

- Improves all cross-validation f-scores
- Combination of PS, TS, SF, SR gives highest f-scores except in RT

### Cross-validation F-scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Stroke category</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>RT</td>
</tr>
<tr>
<td>No aug.</td>
<td>86.7</td>
<td>84.5</td>
</tr>
<tr>
<td>Pitch-shift</td>
<td>87.2</td>
<td>85.5</td>
</tr>
<tr>
<td>Time-scale</td>
<td>88.2</td>
<td>85.0</td>
</tr>
<tr>
<td>Attack-remix</td>
<td>84.3</td>
<td>84.2</td>
</tr>
<tr>
<td>SF-bass</td>
<td>84.5</td>
<td>80.9</td>
</tr>
<tr>
<td>SF-treble</td>
<td>85.8</td>
<td>81.7</td>
</tr>
<tr>
<td>SF-tilt</td>
<td>86.3</td>
<td>82.7</td>
</tr>
<tr>
<td>SF-all</td>
<td>87.6</td>
<td>84.6</td>
</tr>
<tr>
<td>SR-bass</td>
<td>86.0</td>
<td>84.8</td>
</tr>
<tr>
<td>SR-treble</td>
<td>86.1</td>
<td>84.8</td>
</tr>
<tr>
<td>SR-damp.</td>
<td>86.2</td>
<td>85.3</td>
</tr>
<tr>
<td>SR-all</td>
<td>86.8</td>
<td>85.3</td>
</tr>
<tr>
<td>Combined</td>
<td><strong>88.5</strong></td>
<td><strong>84.2</strong></td>
</tr>
</tbody>
</table>
Overall comparison – CV and Test

- Re-trained 1-way CNNs – overall best-performing system (CV and test)
- Data augmentation further improves it, except in test set RB
- Fine-tuned drums CNN gives highest test set RB f-score

<table>
<thead>
<tr>
<th>Method</th>
<th>Stroke Category</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>RT</td>
</tr>
<tr>
<td>RF baseline</td>
<td>86.2</td>
<td>74.2</td>
</tr>
<tr>
<td>3-way drums CNN</td>
<td>74.8</td>
<td>65.4</td>
</tr>
<tr>
<td>1-way CNNs</td>
<td>86.0</td>
<td>79.5</td>
</tr>
<tr>
<td>+ Data-aug</td>
<td><strong>88.5</strong></td>
<td><strong>83.3</strong></td>
</tr>
</tbody>
</table>

Take-aways

- Addressed 4-way tabla stroke classification into musicologically relevant categories
- Introduced diverse, realistic dataset by building on existing ones
- Showed promising results using different approaches that can be brought together
Thank you for your attention!

Questions?