Automatic Labelling of tabla signals

Olivier K. GILLET, Gaël RICHARD
Introduction

- Exponential growth of available digital information
  ⇒ need for Indexing and Retrieval technique

- For musical signals, a transcription would include:
  - Descriptors such as genre, style, instruments of a piece
  - Descriptors such as beat, note, chords, nuances, etc…
  


    - Most effort on isolated sounds

    - Almost no effort on non-Western instrument recognition

- OBJECTIVE : Automatic transcription of real performances of an Indian instrument: the tabla
Outline

- Introduction
- Presentation of the tabla
- Transcription of tabla phrases
  - Architecture of the system
  - Features extraction
  - Learning and classification
- Experimental results
  - Database and evaluation protocols
  - Results
- Tablascope: a fully integrated environment
  - Description & applications
  - Demonstration
- Conclusion
Presentation of the tabla

- **The tabla**: an percussive instrument played in Indian classical and semi-classical music

*The Dayan*: wooden treble drum played by the right hand

*The Bayan*: metallic bass drum played by the left hand
Presentation of the tabla (2)

- Musical tradition in India is mostly oral
  - Use of mnemonic syllables (or *bol*) for each stroke

- Common bols:
  - *Ge, Ke* (bayan bols), *Na, Tin, Tun, Ti, Te* (dayan bols)
  - *Dha* (Na+Ge), *Dhin* (Tin + Ge), *Dhun* (Tun + Ge)

- Some specificities of this notation system
  - Different bols may sound very similar (ex. Ti and Te)
  - Existence of « words »: « TiReKiTe or « GeReNaGe »
  - A mnemonic may change depending on the context
  - Complex rhythmic structure based on *Matra* (i.e main beat), *Vibhag* (i.e measure) and *avartan* (i.e phrase)
Presentation of tabla (3)

- In summary:
  - A tabla phrase is then composed of successive bols of different duration (note, half note, quarter note) embedded in a rhythmic structure
  - Grouping characteristics (words): similarity with spoken and written languages: Interest of « Language models » or sequence models

- In this study, the transcription is limited to
  - the recognition of successives bols
  - The relative duration (note, half note, quarter note) of each bol.
Transcription of tabla phrases

- Architecture of the system

Audio file \(\rightarrow\) Onset detection \(\rightarrow\) Tempo detection \(\rightarrow\) Features extraction \(\rightarrow\) HMM model

Realtime audio input \(\rightarrow\) Phrase correction

Transcription

User input \(\rightarrow\) Phrase correction

Audio output

Synthesis

Annotated corpus

Transcription system

Additional features

64 phrases

5715 bols
Parametric representation

- **Segmentation in strokes**
  - Extraction of a low frequency envelope (sampled at 220.5 Hz)
  - Simple Onset detection based on the difference between two successives samples of the envelope.

- **Tempo extraction**
  - Estimated as the maximum of the autocorrelation function of the envelope signal in the range {60 – 240 bpm}
Features extraction

\[
Dha = Ge + Na
\]
Features extraction

- 4 frequency bands
  - B1 = [0 – 150] Hz
  - B2 = [150 – 220] Hz
  - B3 = [220 – 380] Hz
  - B4 = [700 – 900] Hz

- In the case of single mixture, each band is modelled by a Gaussian

  Feature vector $F = f_1..f_{12}$ (mean, variance and relative weight of each of the 4 Gaussians)
Learning and Classification of bols

- 4 classification techniques were used.
  - K-nearest Neighbors (k-NN)
  - Naive Bayes
  - Kernel density estimator
  - HMM sequence modelling
Learning and Classification of bols

- Context-dependant models (HMM)

```
x(1)  x(2)  x(3)  x(4)  x(5)  x(6)

. . Dha Tin Na Tin Na Ke

```

```
0 1 2 3 4 5

. . . . . .

Dha DhaTin NaTin NaKe

```

```
\begin{align*}
x(3) & \quad x(5) \\
\end{align*}
```

TinNa

NaKe
Learning and Classification of bols

- **Hidden Markov Models**

  - **States:** a couple of Bols $B_1B_2$ is associated to each state $q_t$
  
  - **Transitions:** if state $i$ is labelled by $B_1B_2$ and $j$ by $B_2B_3$ then the transition from state $q_t = j$ to state $q_{t-1} = i$ is given by:
    
    $$a_{i,j} = p(q_t = j | q_{t-1} = i)$$
    
    $$= p(b_t = B_3 | b_{t-1} = B_2, b_{t-2} = B_1)$$

  - **Emissions probabilities:** Each state $i$ labelled by $B_1B_2$ emits a feature vector according to a distribution $b_i(x)$ characteristics of the bol $B_2$ preceded by $B_1$
    
    $$b_i(x) = p(O_t = x | q_t = i)$$
    
    $$= p(O_t = x | b_t = B_2, b_{t-1} = B_1)$$
Learning and Classification of bols

- **Training**
  - Transition probabilities are estimated by counting occurrences in the training database
  - Emission probabilities are estimated with
    - mean and variance estimators on the set of feature vectors in the case of simple Gaussian model
    - 8 iterations of the Expectation-Maximisation (EM) algorithm in the case of a mixture model

- **Recognition**
  - Performed using the traditionnal Viterbi algorithm
Experimental results

- **Database**
  - 64 phrases with a total of 5715 bols
  - A mix of long compositions with themes / variations (*kaïda*), shorter pieces (*kudra*) and basic *taals*.
  - 3 specific sets corresponding to three different tablas:

<table>
<thead>
<tr>
<th></th>
<th>Tabla quality</th>
<th>Dayan tuning</th>
<th>Recording quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabla #1</td>
<td>Low (cheap)</td>
<td>in C#3</td>
<td>Studio equipment</td>
</tr>
<tr>
<td>Tabla #2</td>
<td>High</td>
<td>In D3</td>
<td>Studio equipment</td>
</tr>
<tr>
<td>Tabla #3</td>
<td>High</td>
<td>In D3</td>
<td>Noisier environment</td>
</tr>
</tbody>
</table>
Evaluation protocols

- **Protocol #1:**
  - Cross-validation procedure
    - Database split into 10 subsets (randomly selected)
    - 9 subsets for training, 1 subset for testing
    - Iteration by rotating the 10 subsets
    - Results are average of the 10 runs

- **Protocol #2:**
  - Training database consists of 100% of 2 sets
  - Test is 100% of the remaining sets

  ➤ Different instruments and/or conditions are used for training and testing
### Experimental results (protocol #1)

<table>
<thead>
<tr>
<th>Database # of <em>bols</em></th>
<th>All 5715</th>
<th>Tabla #1 1678</th>
<th>Tabla #2 2216</th>
<th>Tabla #3 1821</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification using only features of stroke $n$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel density estimator</td>
<td>81.7%</td>
<td>81.8%</td>
<td>82.4%</td>
<td>85.2%</td>
</tr>
<tr>
<td>5-NN</td>
<td>83.0%</td>
<td>81.7%</td>
<td>83.3%</td>
<td>85.6%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>76.6%</td>
<td>79.4%</td>
<td>78.6%</td>
<td>78.5%</td>
</tr>
<tr>
<td><strong>Classification using features of stroke $n, n-1, n-2$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel density estimator</td>
<td>86.8%</td>
<td>86.0%</td>
<td>88.7%</td>
<td>92.0%</td>
</tr>
<tr>
<td>5-NN</td>
<td>88.9%</td>
<td>87.2%</td>
<td>88.4%</td>
<td>90.6%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>81.8%</td>
<td>86.5%</td>
<td>83.8%</td>
<td>85.8%</td>
</tr>
<tr>
<td><strong>Classification using language modelling</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM, 3-grams, 1 mixture</td>
<td>88.0%</td>
<td>90.6%</td>
<td>89.9%</td>
<td>92.6%</td>
</tr>
<tr>
<td>HMM, 4-grams, 2 mixtures</td>
<td>93.6%</td>
<td>92.0%</td>
<td>91.9%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>
## Experimental results (protocol #2)

<table>
<thead>
<tr>
<th>Training set</th>
<th>Tabla #1 &amp; Tabla #2</th>
<th>Tabla #2 &amp; Tabla #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set</td>
<td>Tabla #3 (noisy rec.)</td>
<td>Tabla #1 (cheap quality)</td>
</tr>
<tr>
<td>5-NN</td>
<td>79.8 %</td>
<td>78.2 %</td>
</tr>
<tr>
<td>HMM, 3-grams, 1 mixture</td>
<td>90.2 %</td>
<td>88.4 %</td>
</tr>
<tr>
<td>HMM, 4-grams, 2 mixtures</td>
<td>84.5 %</td>
<td>85.0 %</td>
</tr>
</tbody>
</table>

- HMM approaches are more robust to variability
- Simpler classifiers fail to generalise and to adapt to different recording conditions or instruments
## Experimental results

### Confusion matrix by bol category

*HMM 4-grams, 2 mixture classifier*

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1241</td>
<td>22</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>a: resonant <em>dayan</em> strokes (Tin, Na, Tun...)</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1076</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>b: <em>bayan</em>+<em>dayan</em> strokes (Dhin, Dha...)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>766</td>
<td>5</td>
<td>20</td>
<td>c: resonant <em>bayan</em> strokes (Ge, Gi...)</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>448</td>
<td>61</td>
<td>d: non-resonant <em>bayan</em> strokes (Ke, Ki...)</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>50</td>
<td>1938</td>
<td>e: non-resonant <em>dayan</em> strokes (Te, Ti, Tek...)</td>
</tr>
</tbody>
</table>
Tablascope: a fully integrated environment

Applications:
- Tabla transcription
- Tabla sequence synthesis
- Tabla-controlled synthesizer
Conclusion

- A system for automatic labelling of tabla signals was presented
- Low error rate for transcription (6.5%)
- Several applications were integrated in a friendly environment called Tablascope.
- This work can be generalised to other types of percussive instruments

...still need a larger database to confirm the results.....