MIREX SUBMISSION FOR DRUM TRANSCRIPTION 2018

Richard Vogl1,2  Peter Knees1
1 Faculty of Informatics, Vienna University of Technology, Austria  
2 Institute of Computational Perception, Johannes Kepler University Linz, Austria  
{richard.vogl, peter.knees}@tuwien.ac.at

ABSTRACT

This extended abstract provides an overview of the algorithms submitted for the 2018 MIREX drum transcription task. The implemented algorithms are available as a separate build of the madmom framework [1] including the models trained on the public MIREX drum transcription train data http://ifs.tuwien.ac.at/~vogl/mirex2018/.

1. SUBMISSION

The algorithms submitted are modifications of the methods presented in [4] and [6]. For the 3-instrument-class task the same models as submitted in 2017 are used [5]. In the next sections only the modifications for the MIREX submissions compared to [4] and [6] are pointed out. For a detailed discussion of architecture and training refer to the original works.

The final models use an ensemble of models trained on the single splits of the public training sets as described in 2.

1.1 RV1

The RV1 submission consists of a convolutional recurrent neural network (CRNN) trained on the 3-instrument-class data. It is made up of using two convolutional layers featuring 32 3x3 filters with batch normalization [3], followed by a 3x3 max pooling layer. After that another two convolutional layers with 64 3x3 filters and another 3x3 max pooling layer are used. Three bidirectional recurrent layers consisting of 60 GRUs [2] each, follow. The output layer consists of three sigmoid nodes, providing activation functions for each instrument under observation.

Compared to [4] we do not use difference spectrogram as additional input features, since the CNN layers are able to perform the difference calculation easily. Additionally a frequency range from 30 to 15,000 Hz is used resulting in an input vector of length 79. This enables the use of valid convolutions without ending up with non-integer sized shapes, which made integration of the models into the madmom framework easier.

1.2 RV2

The RV2 submission consists of a convolutional neural network (CNN) trained on the 3-instrument-class data. It consists of the same building blocks as the CRNN for RV1 but features two dense layers consisting of 250 ReLU units each, instead of the recurrent layers. The same input features as for the CRNN are used.

1.3 RV3

The RV3 submission is represented by a CRNN trained on the 8-instrument class data. It consists of two convolutional layers using 32 3x3 filters, followed by a 1x3 max pooling layer and another two convolutional layers with 64 3x3 filters followed by another 1x3 max pooling layer. Again all convolutional layers feature with batch normalization. Three bidirectional recurrent layers consisting of 50 GRUs each, follow after the convolutional layers. The output layer consists of eight sigmoid nodes, again representing the eight instrument classes under observation.

As input features the same features as for RV1 and RV2 are used.

Figure 1 visualizes the different architectures used for the individual submissions.

2. TRAINING

Models RV1 and RV2 are trained the same way as described in [4]. A four-split cross-validation training was
Table 1. Overall F-measure results and results on individual sub sets of the MIREX’18 drum transcription task for the submitted models. The two values provided in each column represent mean and sum F-measure values, respectively.

<table>
<thead>
<tr>
<th></th>
<th>RV1</th>
<th>RV2</th>
<th>RV3</th>
</tr>
</thead>
<tbody>
<tr>
<td>overall</td>
<td>0.69 / 0.74</td>
<td>0.65 / 0.72</td>
<td></td>
</tr>
<tr>
<td>IDMT</td>
<td>0.66 / 0.72</td>
<td>0.66 / 0.73</td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td>0.65 / 0.68</td>
<td>0.63 / 0.67</td>
<td></td>
</tr>
<tr>
<td>MEDLEY</td>
<td>0.64 / 0.65</td>
<td>0.55 / 0.55</td>
<td></td>
</tr>
<tr>
<td>RBMA</td>
<td>0.72 / 0.74</td>
<td>0.70 / 0.74</td>
<td></td>
</tr>
<tr>
<td>GEN</td>
<td>0.78 / 0.81</td>
<td>0.70 / 0.75</td>
<td></td>
</tr>
</tbody>
</table>

3. RESULTS

The mean and sum F-measure results for the overall evaluation as well as the sub data sets are shown in this abstract—c.f. Table 1. Overall task winning results are highlighted using bold letters. For the full results consult the results page of the task on the MIREX website.

4. ACKNOWLEDGEMENTS

This work has been partly funded by the Austrian FFG under the BRIDGE 1 project SmarterJam (858514).

5. REFERENCES


