STRUCTURAL SEGMENTATION WITH CONVOLUTIONAL NEURAL NETWORKS MIREX SUBMISSION

Jan Schlüter **Karen Ullrich** **Thomas Grill**
Austrian Research Institute for Artificial Intelligence, Vienna
{jan.schlueter,karen.ullrich,thomas.grill}@ofai.at

ABSTRACT

This submission to the MIREX 2014 Music Structural Segmentation task employs a Convolutional Neural Network (CNN) to detect likely positions of structural boundaries in an audio signal. The network was trained on mel-scaled spectrograms with human structural annotations following the SALAMI guidelines. Our submission solely tackles the task of finding boundaries; it does not attempt to assign labels to detected segments. It is based on our work recently presented in Ullrich et al. [3], using a slightly improved network architecture. This, in turn, is an adaptation of our work in Schlüter et al. [1], the highest-performing submission to the MIREX 2013 Audio Onset Detection task.

1. INTRODUCTION

In order to detect structural segment boundaries in digital audio, we use an artificial neural network trained in a supervised fashion on human-annotated data. For this, we formulate boundary prediction as a binary classification problem: Given an excerpt of an audio signal, decide whether there is a structural boundary at its center or not. Once we have a model solving this problem, we can apply it to a sequence of excerpts extracted in a sliding-window fashion to obtain a curve of boundary probabilities. We search for peaks in this curve in order to predict boundaries in the given music piece.

Here, the music excerpts are represented as mel-scaled spectrograms, the classifier is a Convolutional Neural Network (CNN), and the human-annotated data is an excerpt of the public SALAMI dataset [2] plus additional data annotated according to the same guidelines. The training data was carefully selected to be disjoint from the three datasets used in the MIREX evaluation campaign.

In [3], our method achieved results considerably outperforming any submission to MIREX 2012 and 2013 on a subset of the SALAMI dataset, which contains both classical and popular music recorded under studio conditions and in live concerts. For MIREX 2014, we submit the two best-performing neural networks of [3] (for a time tolerance of ± 0.5 s and ± 3 s, respectively), only slightly modifying their architecture to increase the model capacity.

2. METHOD

As our method is detailed in [3], here we will only give an overview and point out what has changed compared to the previously published work.

2.1 Feature Extraction

From the audio signal, we compute a mel-scaled logarithmic-magnitude spectrogram of 80 bands. To be able to train and predict on spectrogram excerpts near the beginning and end of a music piece, we pad the spectrogram with low-volume pink noise. To allow the CNN to process large temporal contexts while keeping the input size small, we subsample the spectrogram by taking the maximum over adjacent time frames (without overlap). Specifically, for this submission, we either use 16 s or 32 s long spectrogram excerpts subsampled to 116 frames (a frame rate of 7.18 fps or 3.59 fps, respectively). For details on the spectrogram and mel bank parameterization, please refer to [3, Sec. 3.1].

2.2 Network Architecture

As described in the previous section, the input to our CNNs is a spectrogram excerpt of 116 frames and 80 mel bands. The first network layer is convolutional, with 32 filters of 8 frames by 6 bands, followed by max-pooling over patches of 3 frames and 6 bands, another convolutional layer with 64 filters of 6 frames by 3 bands, a fully-connected layer of 128 units and a fully-connected output unit. Both convolutional layers use $\tanh$ units, the fully-connected layers use logistic sigmoid units. This architecture is identical to the one described in [3, Sec. 3.2], except that the number of filters in the two convolutional layers was doubled, slightly improving results on our validation set.

2.3 Network Training

Our networks are trained and validated on a set of 733 music pieces annotated to the SALAMI guidelines, but disjoint from the three datasets used in the MIREX Music Structural Segmentation evaluation [3, Sec. 4]. We used 633 pieces for training, and 100 pieces for validation, to

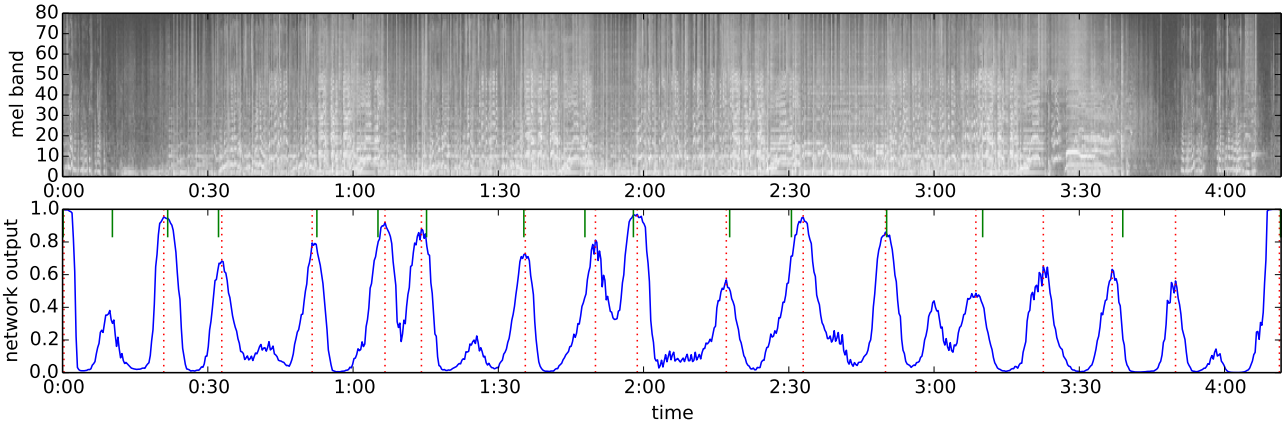

Figure 1. Example spectrogram and network output (*The Weight* by *Rachel Weber*, SALAMI id 1304). For every spectrogram time frame, the network computes an output value. Concatenating all values, we obtain a curve of boundary probabilities for the entire music piece (blue). Local maxima of this curve are boundary candidates, and thresholding them selects the boundary predictions (red, dotted). Ground-truth annotations are shown as short vertical bars (green).

find the best-performing configurations both for our study in [3] and for our MIREX submission.

As in our previous work on CNN-based onset detection [1], the networks are trained on pairs of spectrogram excerpts and binary labels. To cope with the higher temporal inaccuracy and scarcity of boundary annotations compared to musical onset annotations, we employ a technique we termed *target smearing*, detailed in [3, Sec. 3.3]. Specifically, for this submission, the network trained on 16 s excerpts use a smearing width of 1.5 s, and the network trained on 32 s excerpts use a smearing width of 6 s.

Training is performed with mini-batched gradient descent and dropout. For the exact parameters, please see [3, Sec. 3.3].

2.4 Boundary prediction from network output

After training, the networks are applied to pieces of music. For every spectrogram excerpt, the network computes a scalar output between 0 and 1, which can be interpreted as the probability of a boundary occurring at the center of the excerpt. By applying the network to a sequence of excerpts, advancing a single time frame between each, we obtain a curve for the entire music piece (this can be efficiently implemented as a series of convolutions and a final dot product). With peak-picking and thresholding, we obtain boundary locations from this curve. The peak-picking threshold for a given network is chosen to optimize the boundary retrieval F-score on the validation set. An example for a network output curve, a set of predicted and annotated boundaries is given in Figure 1. For details on the peak-picking algorithm, please see [3, Sect. 3.4].

For improved results, we train five identically-parameterized networks, starting from different random weight initializations, and average their output before peak-picking. (This is a standard technique known as *bagging*.)

2.5 Difference of our two submissions

We submit two variants to MIREX 2014: SUG1 and SUG2. The former is trained on 16 s spectrogram excerpts at 7.18 frames per second, using a target smearing width of 1.5 s. This model achieved the highest boundary retrieval F-score on our validation set for a retrieval tolerance of ± 0.5 s, using a threshold of 0.37. The second variant, SUG2, is trained on 32 s excerpts at 3.59 fps at a smearing width of 6 s. It was the best-performing model for a retrieval tolerance of ± 3 s, using a somewhat lower threshold of 0.21 (returning more boundaries).

3. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF): TRP 307-N23.

4. REFERENCES

[1] J. Schlüter and S. Böck. Improved musical onset detection with convolutional neural networks. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[2] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 555–560, 2011.

[3] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary Detection in Music Structure Analysis using Convolutional Neural Networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.