# MUSIC BOUNDARY DETECTION USING NEURAL NETWORKS ON COMBINED FEATURES AND TWO-LEVEL ANNOTATIONS

**Thomas Grill and Jan Schlüter**
Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
thomas.grill@ofai.at, jan.schlueter@ofai.at

## ABSTRACT

The determination of structural boundaries is a key task for understanding the structure of a musical piece, but it is also highly ambiguous. Recently, Convolutional Neural Networks (CNN) trained on spectrogram features and human annotations have been successfully used to tackle the problem, but still fall clearly behind human performance. We expand on the CNN approach by combining spectrograms with self-similarity lag matrices as audio features, thereby capturing more facets of the underlying structural information. Furthermore, in order to consider the hierarchical nature of structural organization, we explore different strategies to learn from the two-level annotations of main and secondary boundaries available in the SALAMI structural annotation dataset. We show that both measures improve boundary recognition performance, resulting in a significant improvement over the previous state of the art. As a side-effect, our algorithm can predict boundaries on two different structural levels, equivalent to the training data.

## 1. INTRODUCTION

The decomposition of a piece of music into parts known as movements, phrases, chorus and verse, etc., also commonly referred to as *musical form*, is an important task and a major challenge in music analysis. However, the identification and exact placement of transition points, or, *boundaries* between such structural elements is often indistinct, even for trained human annotators. Figure 1 represents an excerpt of the piece "The Wet Spot" by "Southern Culture On The Skids" (index 1358 in the SALAMI collection, see Section 4.1). Two different sets of human-annotated boundaries (*ground truth*) are depicted by vertical marks at the top and bottom of the plots. They clearly illustrate the ambiguity of annotating boundaries at a certain level of detail. The annotators agreed well on the positions of the boundaries, but for some of these they disagreed whether they should be considered strong (or 'coarse', delimiting 'large scale', resp., 'functional' sections) [1] or weak ('fine', delimiting 'small scale' sections). This poses a problem as the common methodology used for the evaluation of structural annotation ignores the hierarchical nature and considers only one level of detail, usually the coarse boundaries.

The currently by far best-performing methods for boundary detection use Convolutional Neural Networks (CNNs), trained on large corpora of human-annotated structural annotations. The algorithms are based on mel-scaled log-magnitude spectrograms (MLSs), taking into account a relatively short context of a few seconds, depending on the desired precision. As shown in Figure 1a, the CNN based solely on an MLS or a variation such as MLS-HPSS (Harmonic-Percussive Source Separation, see [1]), has difficulties of identifying certain boundaries, indicated by low probabilities in the prediction curve (Figure 1b). We have investigated in [3] that *self-similarity lag matrices* (SSLMs, see Figures 1c and 1d) can be used as additional alternative structural information to significantly improve boundary detection.

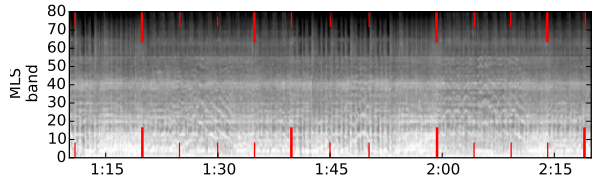
In this contribution, we expand on our approach by combining more input features, and put particular focus on the integration of multiple and two-level annotation ground-truth, as available in the SALAMI dataset. The structure of the paper is as follows: After giving an overview over related work in Section 2, we propose our method in Section 3. In Section 4, we describe the experimental setup and our evaluation strategy. Section 5 presents our main results. We wrap up in Section 6 with a discussion and outlook.

## 2. RELATED WORK

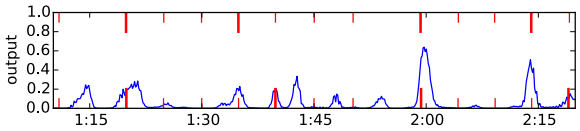
Following the overview paper by Paulus et al. [12], three fundamental approaches to music structure analysis can be distinguished: Novelty-based, detecting transitions between contrasting parts, homogeneity-based, identifying sections that are consistent with respect to their musical properties, and repetition-based, building on the determination of recurring patterns. Novelty is typically computed from self-similarity matrices (SSMs) or self-distance matrices (SDMs) by sliding a checkerboard kernel along the diagonal [2], building on audio descriptors like MFCCs, pitch class profiles, or rhythmic features [10]. Turnbull



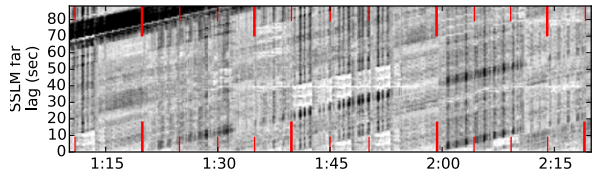
[1] See [16] and SALAMI Annotator's Guide, http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf, accessed 2015-05-04



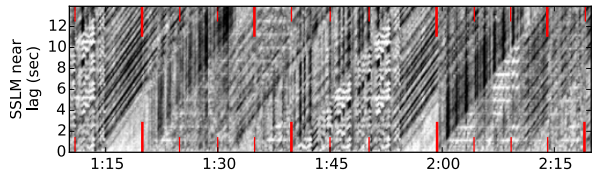

(a) Mel-scaled log-magnitude spectrogram



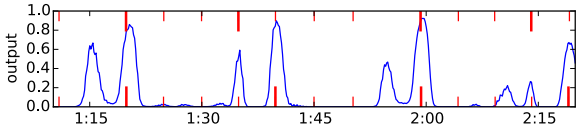

(b) CNN predictions for coarse boundaries on HPSS-decomposed mel spectrogram



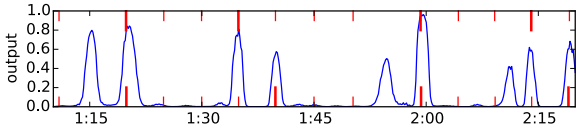

(c) Self-similarity lag matrix, 88 seconds context (SSLM far)



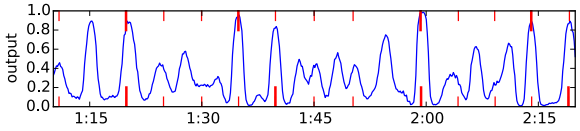

(d) Self-similarity lag matrix, 14 seconds context (SSLM near)




(e) CNN predictions for coarse boundaries on SSLMs near and far combined




(f) CNN predictions for coarse boundaries on MLS-HPSS and SSLMs near and far combined




(g) CNN predictions for fine boundaries on MLS-HPSS and SSLMs near and far combined

**Figure 1**: Boundary recognition using CNNs on different underlying audio features, illustrated on the piece "The Wet Spot" by "Southern Culture On The Skids". Two sets of human annotation ground-truth are shown in red on top and bottom of each plot. Coarse boundaries are thick, fine boundaries are thin. Visit `http://www.ofai.at/research/impml/projects/audiostreams/ismir2015` for a version with audio.

et al. [17] compute difference features on more complex audio feature sets and use trained Boosted Decision Stumps for boundary detection. In order to capitalize on repeated patterns, SSMs or SDMs are used with various heuristic rules and optimization schemes for structure formation [4, 9, 11]. McFee and Ellis employ spectral clustering [6], or add a supervised learning scheme using ordinal linear discriminant analysis and constrained clustering [5]. When using end-to-end neural network techniques such as Ullrich et al.'s CNNs [18], the separation between the fundamental approaches becomes blurred as the CNN infers the relationships between audio features and ground truth from the provided training data. In a similarly integral fashion, Serrà et al. [15] propose an unsupervised method explicitly combining all three domains.

## 3. PROPOSED METHOD

Our approach is derived from the work by Ullrich et al. [18]. In the following, we will mainly describe our extensions to this method.

### 3.1 Feature extraction

For each audio file under analysis, we first compute a STFT magnitude spectrogram with a window size of 46 ms (2048 samples at 44.1 kHz sample rate) and 50% overlap, and apply a mel-scaled filterbank of $n = 80$ triangular filters from 80 Hz to 16 kHz and scale magnitudes logarithmically.

From this MLS we compute a HPSS decomposition with a kernel size of $21 \times 21$ bins. Preliminary experiments showed that the actual size is a rather insensitive parameter. We either use MLS only or MLS-HPSS (two parallel channels) as one part of the network input.

Our method of generating the SSLMs, which represent similarities of the MLS at one point in time in relation to points in the past, up to a certain *lag time*, is derived from work by Serrà et al. [15] and described in detail in [3]. We use the MLS time series $\mathbf{x}_{i=1\ldots N}$ from above, downsample it by max-pooling of a factor $p = 2$, and apply a DCT-II transformation on each frame with the static component omitted. Several of these frames are concatenated within a local time context of $L$ bins, equivalent to 0.1 seconds, resulting in the time series $\hat{\mathbf{x}}_i$. A cosine distance function $\delta_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle$ is used to build the $\lfloor \frac{N}{p} \rfloor \times \lfloor \frac{L}{p} \rfloor$ recurrence matrix

$$D_{i,l} = \delta_{\cos}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_{i-l}), \quad l = 1 \ldots \lfloor \tfrac{L}{p} \rfloor. \tag{1}$$

To reveal relationships between distances across this matrix, adaptive thresholding is performed with a smooth sigmoid transfer function $\sigma(x) = 1/(1 + e^{-x})$, yielding

$$R_{i,l} = \sigma\left(1 - \frac{D_{i,l}}{\varepsilon_{i,l}}\right). \tag{2}$$

The adaptive threshold, or, in this context, equalization factor $\varepsilon_{i,l}$ is set to a quantile $Q_\kappa$ with $\kappa = 0.1$ of the distances $\delta_{\cos}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_{i-j})$ and $\delta_{\cos}(\hat{\mathbf{x}}_{i-l}, \hat{\mathbf{x}}_{i-l-j})$ for $j = 1 \ldots \lfloor \frac{L}{p} \rfloor$,

or

$$\varepsilon_{i,l} = Q_\kappa \left( D_{i,1}, \ldots, D_{i,\lfloor \frac{L}{p} \rfloor}, D_{i-l,1}, \ldots, D_{i-l,\lfloor \frac{L}{p} \rfloor} \right). \quad (3)$$

All indices $i < 1$ are wrapped around to $i' = i + \lfloor \frac{N}{p} \rfloor$, resulting in a time-circular SSLM.

### 3.2 Feature preprocessing

Like [18], for the MLS features, we pad the spectrogram with pink noise of $-70$ dB FS as needed to process the beginning and end of a piece. For the MLS-HPSS variant, the harmonic and percussive components are separated at this point. After subsampling the MLS by taking the maximum over 6 adjacent time frames without overlap (max-pooling), we normalize to zero mean and unit variance for each frequency band. For the SSLM features, we use circular padding and pooling factors examined in [3]: A factor of 3 for a time context of 14 seconds (feature 'SSLM-near'), and a factor of 19 for a context of 88 seconds (feature 'SSLM-far'). We then also normalize each lag band to zero mean and unit variance.

### 3.3 Convolutional neural network

CNNs are feed-forward networks that include *convolutional layers* computing a convolution of their input with small learned filter kernels of a given size. This allows processing large inputs with few trainable parameters, and retains the input's spatial layout. When used for binary classification, the network usually ends in one or more dense layers integrating information over the full input at once, discarding the spatial layout. Our architecture for this work is based on the one used by Ullrich et al. [18] on MLS features for their MIREX submission [14]. It has a convolutional layer of 32 $8 \times 6$ kernels (8 time frames and 6 frequency bands), a max-pooling layer of $3 \times 6$, another convolution of 64 $6 \times 3$ kernels, a dense layer of 128 units and a dense output layer of 1 unit.

We employ a variant of this architecture to support multiple input features instead of one. A comparison of different architectural variations has been shown in [3], where a late 'time-synchronous fusion' of the input features, performed in the last convolutional layer, yielded the best results: since the input features cover the same temporal context at the same resolution, their feature maps can be synchronously convolved over time. Figure 2 shows the underlying CNN architecture used for all experiments in our study. The inputs (bottom) are varied, e.g., MLS only is used instead of MLS-HPSS, or one of the input legs is left out. For the outputs (top), either only the coarse unit is used, or both coarse and fine.

Training is done by mini-batch gradient descent, using the same hyper-parameters and tweaks as Ullrich et al. [18]. Likewise, we follow the peak-picking strategy described therein to retrieve likely boundary locations from the network output.

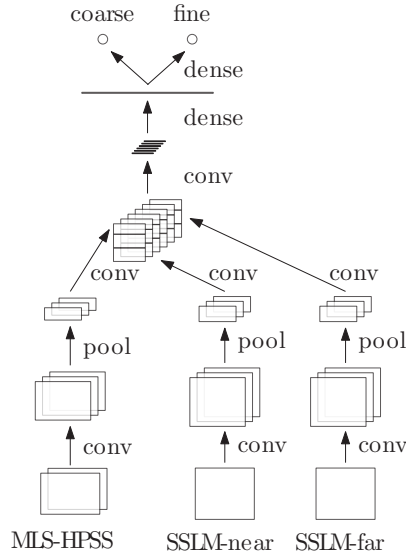



**Figure 2**: The CNN architecture in use for all the models. The full model is shown here, inputs or outputs were varied for the different experiments.

## 4. EXPERIMENTS

### 4.1 Data set

We base our experiments on the data set described by Ullrich et al. [18] which is a subset of the Structural Analysis of Large Amounts of Music Information (SALAMI) [16] version 1.2 database. A part of this SALAMI 1.2 data set was also used in the "Audio Structural Segmentation" task of the annual MIREX evaluation campaign in the years 2012 through 2014. [2] Lately, the data set has been updated to version 2.0 [3] with a large number of issues fixed. The entire data set contains over 1600 musical recordings of different genres and origins. In SALAMI version 2.0, a total of 1164 recordings (with 763 double-annotated) are publicly available. Identically to [18], we used 633 musical pieces for training, 100 for validation and 487 pieces as a test set for final evaluation of our models against the published results of the various MIREX submissions.

### 4.2 Evaluation

For the MIREX campaign's boundary retrieval task, three different evaluation measures are used: *Hit rate* for time tolerances $\pm 0.5$ and $\pm 3$ seconds, and *Median deviation*. The latter computes the median time distance between each annotated boundary and its closest predicted boundary, and vice versa. The former checks which predicted boundaries fall close enough to an unmatched annotated boundary (true positives), records remaining unmatched predictions and annotations as false positives and negatives, respectively, and computes the precision, recall and $F_1$ scores. The Hit rate $F_1$ score is the measure most frequently used in the literature.

[2] Music Information Retrieval Evaluation eXchange, `http://www.music-ir.org/mirex`, accessed 2015-04-30

[3] `https://github.com/DDMAL/salami-data-public/releases/tag/2.0`, accessed 2015-04-30

As explicated in [18], baseline scores can be estimated using variations of regularly or randomly spaced grids as synthetic boundary estimates. For an evaluation tolerance of $\pm 0.5$ seconds, the baseline within our test data set is $F_1 \approx 0.15$. Upper bounds, on the other hand, can be derived from the differences between two independent annotations of the same musical pieces. By analyzing the items within our test data set that have been annotated twice (439 pieces), we calculated $F_1 \approx 0.74$.

In the existing literature, both tolerances of $\pm 0.5$ and $\pm 3$ seconds are commonly used. For this contribution, due to space constraints, we only evaluate for $\pm 0.5$ seconds, where the explorable space, that is, the distance between the lower baseline and the upper bound exhibited in human ground-truth annotations is much greater than for $\pm 3$ seconds (with lower and upper bounds at 0.33 and 0.80, respectively). Our evaluation code is equivalent to the boundary detection implemented in `mir_eval` [13], omitting the borders at the beginning and end of sound files.

Nieto et al. [8] have identified the $F_{0.58}$ measure to be more perceptually informative than the typically used $F_1$ measure. As this is a relatively new finding and it is not as well established as the $F_1$ measure (which is, e.g., used in MIREX), we base threshold optimization and model selection on the latter.

### 4.3 Combination of features

Building on [3], we combine mel-scaled log-magnitude spectrograms (MLS) and self-similarity lag matrices (SSLM) as input features to the CNN. A decomposition of MLS into harmonic and percussive components (feature 'MLS-HPSS') and the combination of two SSLMs, one a high-resolution, low lag matrix, the other one a low-resolution, high lag matrix, provides even more structural information to the network. We mainly compare two models: 'MLS + SSLM-near' (the model developed in [3]), and the more complex and computationally more expensive model 'MLS-HPSS + combined SSLM', integrating all available input features.

The different input features are fused at a relatively late stage in the network (see Figure 2), using a convolutional layer which spans all the vertical (frequency or lag time) components, but only a very short time context. This is motivated by the assumption that the input features are strongly correlated in time. Figure 3 shows boundary recognition scores for the 'MLS + SSLM-near' model and three different context widths (1, 3 and 5 bins), evaluated on the validation set. As can be seen, a temporal context for the fusion layer of more than a single bin does not improve the results.

### 4.4 Consideration of multiple annotations

Up to now, CNN-based boundary recognition algorithms have been trained on data sets with just one annotation version per music piece. SALAMI data contains double annotations for the majority of training examples. It is worth inspecting whether multiple, potentially contradicting annotations help or confuse the CNN training process. Figure 4 shows the results for three different models trained with single and multiple annotations, respectively, evaluated on the validation set. Employing multiple annotations by duplicating audio features and applying the alternative target annotations, the number of training examples increase from 1198707 (with $70317 \times 3$ positive examples) to 1670944 ($98913 \times 3$ positive examples) data points, corresponding to $+39\%$. A positive effect can be observed for models with more versatile structural information available for the network. In these cases, the increase of the $F_1$ score is in the range of 1–2%.

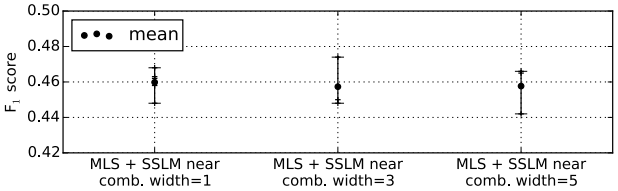



**Figure 3**: Comparison of boundary recognition $F_1$ scores for different widths of the CNN fusion layer.

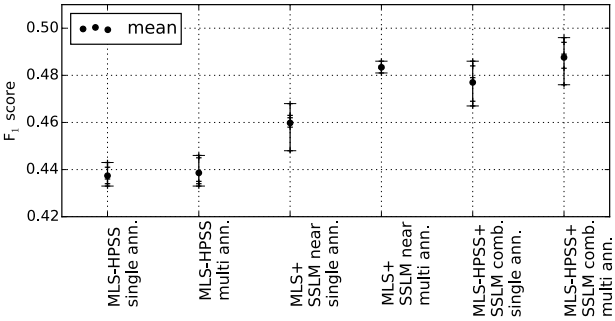



**Figure 4**: Comparison of boundary recognition $F_1$ scores for different models trained with single and multiple annotations.

### 4.5 Integration of fine annotation

Traditionally, boundary detection in MIR has been performed on only one structural level. As motivated in Section 1, we would like to deal with the ambiguity of annotating boundaries at a certain level of detail by capitalizing on the two-level annotations present in the training data set. This way, the neural network should be able to refine its distinction between main and secondary boundaries.

We explored three different modes for the combination of coarse and fine boundaries: Firstly, by using only one target output vector by assigning full training weights to coarse labels and reduced training weights (e.g., factors of 0.3 or 0.5) to fine labels. Secondly, by using two target outputs with equal weights, one for the coarse labels and one for the fine labels ('concat' mode). And finally, using two target outputs, with coarse labels and full weights assigned to the first output vector. Fine labels are assigned to the second output vector, but only where they are distinct from a coarse label ('contrast' mode). This should create a more pronounced contrast between coarse and fine labels

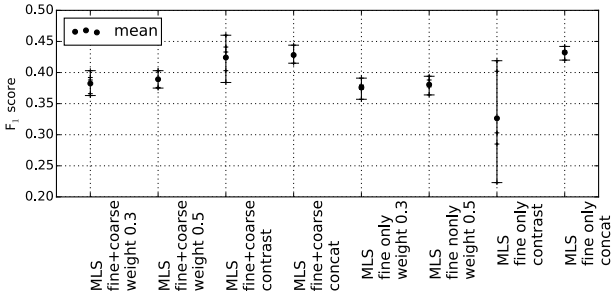



**Figure 5**: Comparison of boundary recognition $F_1$ scores for different integration modes of the second-level 'fine' annotation, evaluated on our validation set.

with the potential danger of some contradiction.

Not for all of our training data two-level annotations were available. We tried two variations: For the first one, we put coarse boundaries where fine ones were not available ('fine + coarse'), and for the second one, we used only those annotations with two levels available ('fine only'), effectively reducing the number of training examples including multiple annotations to 1224891 (with $74400 \times 3$ positive examples).

Figure 5 shows the results for the three combination modes (with different weighting parameters) and two data set variations, computed on MLS input features and evaluated on the validation set. The combination modes for coarse and fine data with two output vectors perform better than the ones with only one output. The 'contrast' mode exhibits instabilities for the results, most probably due to the relatively small validation data set. We selected the best-performing and reliable 'concat' mode with two output units as our working model. The distinction between 'fine + coarse' and 'fine only' variations is more or less inconclusive, with very little advantage for the latter. However, as the spreading of $F_1$ scores is less for 'fine only', we settled for this variation of the 'concat' mode.

## 5. RESULTS

Figure 6 shows boundary recognition scores (on the primary 'coarse' boundaries) of several of our models, with peak-picking thresholds optimized on the validation set, and results evaluated on the test set. Each model variation has been trained and evaluated five times. The individual, mean and 'bagged' results are shown in the graph. 'Bagging' means that the outputs of all five models are averaged and peak-picking is performed on the result, thereby reducing statistical variations. Using a MLS-HPSS decomposition does not score significantly higher than MLS only. Likewise, using a combination of SSLM 'near' (14 seconds lag, high resolution) and 'far' (88 seconds lag, low resolution) does not score higher than SSLM 'near' only. However, in combination, it can be seen that all 'MLS-HPSS + combined SSLM' results are higher than their respective equivalents of 'MLS + SSLM-near'. For both combined models, using multiple annotations raises the scores relative to single annotations. Additional fine annotations for CNN training further increase the scores. On the right-hand-side of Figure 6 different feature combinations using multiple fine annotations are shown. The more perspectives on the audio provided as input, the higher the scores.

| **Algorithm** | $\mathbf{F}_1$ | $\mathbf{F}_{.58}$ | **Rec.** | **Prec.** |
|---|---|---|---|---|
| Upper bound (est.) | .74 | .74 | | |
| *All features, multi+fine ann.* | **.508** | .529 | .502 | .572 |
| *MLS+SSLM-near, multi+fine* | .496 | .506 | .509 | .536 |
| *MLS+SSLM-near, single ann.* | .469 | .466 | .504 | .475 |
| SUG1 (2014) | .422 | .442 | .422 | .490 |
| MP2 (2013) | .294 | .280 | .362 | .271 |
| MP1 (2013) | .276 | .270 | .311 | .269 |
| NB1 (2014) | .270 | .246 | .374 | .229 |
| KSP2 (2012) | .263 | .231 | .422 | .209 |
| Baseline (est.) | .15 | .21 | | |

**Table 1**: Boundary recognition scores for coarse boundaries at a tolerance of $\pm 0.5$ seconds, evaluated on our SALAMI 2.0 test dataset. Comparison of our models (in italics) with the five best-performing algorithms of the MIREX campaigns 2012 through 2014.

| **Algorithm** | $\mathbf{F}_1$ | $\mathbf{F}_{.58}$ | **Rec.** | **Prec.** |
|---|---|---|---|---|
| Upper bound (est.) | .75 | .76 | | |
| *All features, multi+fine ann.* | **.485** | .523 | .443 | .587 |
| *MLS+SSLM-near, multi+fine* | .478 | .515 | .439 | .576 |
| Baseline (est.) | .23 | .17 | | |

**Table 2**: Boundary recognition scores of two of our models for 'fine', second-level boundaries at a tolerance of $\pm 0.5$ seconds, evaluated on our SALAMI 2.0 test dataset.

See Table 1 for a listing of our results in comparison to the best-performing algorithms of the MIREX campaigns 2012 through 2014. All results have been evaluated on SALAMI 2.0 data. Note that the scores are generally lower than for SALAMI 1.2 annotations (cf. [18]). The reason is that in the new data set version many formerly 'trivial boundaries' (sitting at the beginning or end of sound files) have been corrected. These boundaries have moved away from the borders and are now headed, or trailed, respectively, by silence or crowd noise, and are therefore more difficult to predict. The 'MLS+SSLM-near' model trained with single annotations is equivalent to the model used in [3], with an additional dense layer in the present work. 'All features' denotes the 'MLS-HPSS + combined SSLM' model, yielding the best boundary prediction results.

Table 2 lists boundary recognition results of the 'fine' output unit of our network, trained and evaluated on the 'small-scale', second-level annotations of the SALAMI 2.0 data set. To our knowledge, only McFee and Ellis [6] have so far evaluated their algorithms (as well as SMGA [15]) on the secondary boundaries. They report $F_1$ scores up to $0.292 \pm 0.15$ on the SALAMI 1.2 dataset.

Table 3 presents boundary recognition results on the Beatles-ISO dataset,[4] comprised of all 12 Beatles albums with 180 songs in total. We used the best-scoring model from above, using all input features, trained on

[4] `http://isophonics.net/content/reference-annotations-beatles`, accessed 2015-04-30

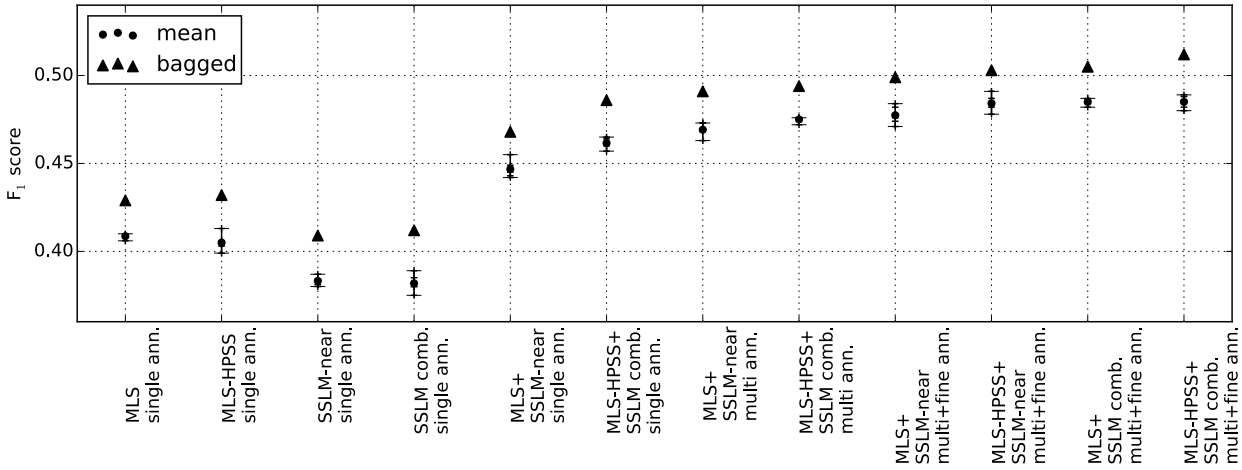



**Figure 6**: Comparison of boundary recognition $F_1$ scores on SALAMI 2.0 data for different models under examination. Threshold optimization performed on validation set, evaluation done on test set.

| Algorithm | $F_1$ | $F_{.58}$ | Rec. | Prec. |
|---|---|---|---|---|
| *All features, multi+fine ann.* | **.558** | .590 | .522 | .640 |
| *MLS+SSLM-near, multi+fine* | .526 | .553 | .500 | .597 |
| SUG1 | .424 | .457 | .385 | .510 |
| MP2-beatles | .334 | .321 | .376 | .311 |
| MP2-salami | .322 | .313 | .355 | .309 |
| NB1 | .286 | .274 | .332 | .266 |
| MP1 | .278 | .280 | .285 | .285 |
| NB2 | .266 | .255 | .302 | .247 |
| NB3 | .227 | .211 | .287 | .200 |
| Baseline (est.) | .15 | .22 | | |

**Table 3**: Boundary recognition scores at a tolerance of $\pm 0.5$ seconds, evaluated on the Beatles-ISO dataset (180 songs). Two of our models are compared to several published state-of-the-art algorithms.

SALAMI 2.0 with multiple coarse and fine annotations. We were able to compare the predictions of our CNN to the best-performing algorithms of last years' MIREX submissions by Schlüter et al. (SUG1, personal communication), McFee and Ellis [5] (MP1 and MP2, the latter optimized either for SALAMI and Beatles data), [5] and Nieto and Bello [7] (NB1, NB2, NB3), [6] respectively. Note that the scores of our models are above those of other state-of-the-art algorithms by a large margin, although we have not trained or tuned our models in any way specifically on the kind of music realized by the Beatles.

## 6. DISCUSSION AND OUTLOOK

In this contribution, we have dealt with the prediction of musically relevant structural boundaries, focused primarily on the stylistically mixed SALAMI data set in its latest version 2.0, with additional evaluation on the Beatles-ISO data set.

We have re-used the CNN architecture developed in [3] with some modifications. On the one hand, we have fed it a combination of different input features and have been able to show that the CNN is able to produce highest-scoring results with HPSS-decomposed mel-scaled spectrograms (MLS) in combination with self-similarity lag matrices (SSLMs) on two different time-scales, covering both structural detail and longer time context. On the other hand, we have taken advantage of the fact that the SALAMI data set is annotated on two structural levels, and, for the most part, by two independent annotators. The integration of this supplementary data helps the CNN to take better informed decisions between primary and secondary boundaries. Evaluated on SALAMI 2.0 data, we have been able to raise the state of the art from the best MIREX submission [14] at $F_1 = 0.422$, and our previous point of reference [3] at $F_1 = 0.469$ to the score of $F_1 = 0.508$ for the best model, integrating all available input features, as well as multiple and two-level annotations. As the CNN model trained on two-level annotation possesses two output units, its subsequent application also yields two independent predictions for 'coarse' and 'fine' boundaries.

Although we have not touched (nor listened to) music by the Beatles while developing our models, evaluation on this data set reveals that our models are quite robust, yielding a boundary recognition score of $F_1 = 0.558$, which is significantly higher than the previously published state of the art.

We are still actively exploring the possibilities of CNNs applied to music structure discovery. That said, we have neither exhaustively researched the space of possible input features, nor all meaningful variations of model architecture and learning parameters. There is plenty of remaining headroom to the 'upper bound' inter-annotator $F_1$ scores.

## 7. ACKNOWLEDGMENTS

This research is funded by the Federal Ministry for Transport, Innovation & Technology (BMVIT) and the Austrian Science Fund (FWF) through project TRP 307-N23 and the Vienna Science and Technology Fund (WWTF) through project MA14-018.

[5] `https://github.com/bmcfee/olda`, accessed 2015-05-01

[6] `https://github.com/urinieto/SegmenterMIREX2014`, accessed 2015-05-01

## 8. REFERENCES


[1] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[2] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'00)*, volume 1, pages 452–455, New York, USA, 2000.

[3] Thomas Grill and Jan Schlüter. Music Boundary Detection Using Neural Networks on Spectrograms and Self-Similarity Lag Matrices. In *Proceedings of the 23rd European Signal Processing Conference (EUSPICO 2015)*, Nice, France, 2015.

[4] Lie Lu, Muyuan Wang, and Hong-Jiang Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 275–282, New York, USA, 2004.

[5] Brian McFee and Daniel P. W. Ellis. Learning to segment songs with ordinal linear discriminant analysis. In *International conference on acoustics, speech and signal processing*, ICASSP, 2014.

[6] Brian McFee and Daniel PW Ellis. Analyzing song structure with spectral clustering. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 405–410, Taipei, Taiwan, 2014.

[7] Oriol Nieto and Juan Pablo Bello. Music Segment Similarity Using 2D-Fourier Magnitude Coefficients. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 664–668, Florence, Italy, 2014.

[8] Oriol Nieto, Morwaread M Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual analysis of the f-measure for evaluating section boundaries in music. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 265–270, Taipei, Taiwan, 2014.

[9] Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *AMCMM '06: Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 59–68, New York, USA, 2006.

[10] Jouni Paulus and Anssi Klapuri. Acoustic features for music piece structure analysis. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[11] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[12] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, pages 625–636, 2010.

[13] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis. mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.

[14] Jan Schlüter, Karen Ullrich, and Thomas Grill. Structural segmentation with convolutional neural networks mirex submission. In *Tenth running of the Music Information Retrieval Evaluation eXchange (MIREX 2014)*, 2014.

[15] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll. Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. In *IEEE Transactions on Multimedia*, 16(5):1229–1240, 2014.

[16] 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 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 555–560, 2011.

[17] Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 51–54, 2007.

[18] 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.