LEARNING MULTIFACETED SELF-SIMILARITY OVER TIME AND FREQUENCY FOR MUSIC STRUCTURE ANALYSIS

Tsung-Ping Chen[1] **and Kazuyoshi Yoshii**[2]
[1]Graduate School of Informatics, Kyoto University, Japan
[2]Graduate School of Engineering, Kyoto University, Japan
chen.tsungping.74e@st.kyoto-u.ac.jp, yoshii.kazuyoshi.3r@kyoto-u.ac.jp

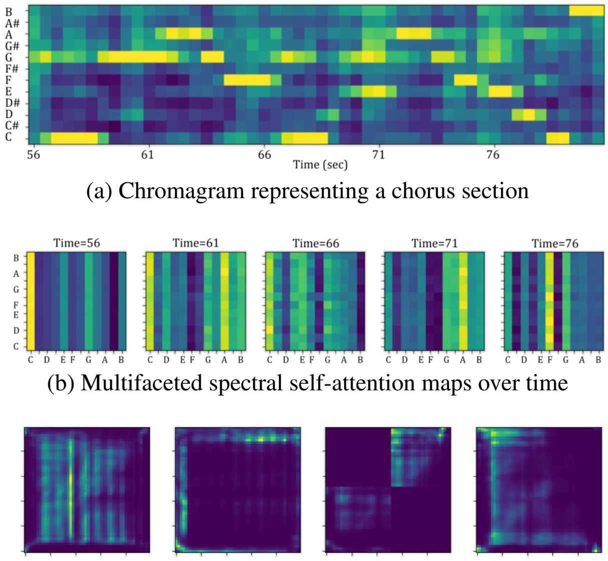
ABSTRACT

This paper describes a deep learning method for music structure analysis (MSA) that aims to split a music signal into temporal segments and assign a function label (e.g., intro, verse, or chorus) to each segment. The computational base for MSA is a spectro-temporal representation of input audio such as the spectrogram, where the compositional relationships of the spectral components provide valuable clues (e.g., chords) to the identification of structural units. However, such implicit features might be vulnerable to local operations such as convolution and pooling operations. In this paper, we hypothesize that the self-attention over the spectral domain as well as the temporal domain plays a key role in tackling MSA. Based on this hypothesis, we propose a novel MSA model built on the Transformer-in-Transformer architecture that alternately stacks spectral and temporal self-attention layers. Experiments with the Beatles, RWC, and SALAMI datasets showed the superiority of the dual-aspect self-attention. In particular, the differentiation between spectral and temporal self-attentions can provide extra performance gain. By analyzing the attention maps, we also demonstrate that self-attention can unfold tonal relationships and the internal structure of music.

1. INTRODUCTION

Music structure refers to the sequential arrangement of musically coherent units that form a musical work. Music structure analysis (MSA) calls for a comprehensive understanding of various musical elements such as rhythm, melody, and harmony, and has still been an open problem in the field of music information retrieval (MIR), partly due to its multifaceted and ill-defined nature [1].

There are two major ways of representing music structure. The *semiotic* representation [2] uses a set of arbitrary symbols (e.g., A-B-C-B-C) for revealing the relationships between segments within a musical piece. The *functional* representation uses a set of semantic labels (e.g., intro-verse-chorus-verse-chorus) for indicating the roles of individual units. The functional representation can be converted into the semiotic one, but not vice versa.



(a) Chromagram representing a chorus section

(b) Multifaceted spectral self-attention maps over time

(c) Multi-scale temporal self-attention maps

Figure 1: Non-local dependencies of music such as chord tones and repeated patterns can be captured with spectral and temporal self-attention mechanisms. (a) The chromagram shows that this musical excerpt is dominated by the C major chord. (b) The spectral attention maps show that pitch classes C, E, and G persistently draw attention while the chroma features vary over time. (c) The temporal attention maps delineate the internal structures of this excerpt.

A common deep learning approach to MSA involves using a convolutional neural network (CNN) to extract latent features from a spectro-temporal representation of input audio, such as a mel spectrogram or chromagram [3–5]. The assumption underlying this approach is the time-frequency locality of musical features, which should be treated with caution when characterizing the global structure of music. Actually, musical elements have non-local dependencies. Chords consist of musical sounds that are widely distributed over frequency. Musical patterns such as chord progressions or musical phrases are commonly repeated over time. Such non-local time-frequency dependencies can hardly be captured by a CNN that aggregates local features while reducing the time-frequency dimensions with pooling operations [6, 7].

One promising architecture for learning non-local dependencies in a music recording is the SpecTNT [8], a variant of the Transformer-in-Transformer (TNT) [9] for modeling spectrogram-like representations. The SpecTNT iterates feature transforms of the multi-head self-attention (MHSA) mechanism [10] alternately along the spectral and temporal axes while keeping the time-frequency dimensions of input features. This method, however, suffers from a potential performance limitation because the individual characteristics of spectral and temporal dimensions are indistinguishable to the MHSA.

To overcome this limitation, we propose to integrate specialized MHSA mechanisms into the SpecTNT architecture for the MSA task regarding the functional representation. As outlined in Figure 1, our method involves gathering spectral and temporal information alternately from an input spectro-temporal representation with two types of MHSA mechanisms. The *spectral self-attention* extracts compositional relationships among two types of spectral features at each time step. The *temporal self-attention* aggregates spectral information at multiple time scales. The proposed method is systematically evaluated with three corpora consisting of popular music. In addition, the attention maps are analyzed to advocate paying attention to the non-locality of musical features.

The main contribution of this work is to emphasize the non-local dependencies of music over frequency as well as that over time. While non-local temporal correlations have been extensively studied, spectral non-locality remains underrepresented. In this concern, we adapt MHSA mechanisms to both aspects and analyze the self-attention maps to elaborate on the non-locality of musical features. This work may draw attention to such delicate characteristics that could be crucial for various tasks in MIR.

2. RELATED WORK

Segmentation and labeling are two subtasks of MSA [11]. The former detects the boundaries of structural units, and the latter categorizes musical segments either by the relationships with the semiotic representation or by the structural roles with the functional representation.

For the segmentation task, a spectro-temporal representation or a sequence of higher-level features extracted by a CNN is typically used to compute a novelty curve [12–16], from which musical boundaries are retrieved with a peak-picking algorithm [17, 18]. A key feature of our method is that we employ CNNs without any pooling layers for feature extraction. Since adjacent spectral beams are irrelevant in the sense of music, naive local pooling would hinder the learning of spectral patterns.

For the labeling task, the similarity-based approach is commonly taken in support of the semiotic representation of music structure [19–24], yet deep learning classification frameworks have recently been introduced for the estimation of structural functions [25, 26]. For both scenarios, the segmentation of an input piece will be a byproduct of the labeling task. However, a smoothing method is typically required to refine the fragmented segmentation results caused by unusual label changes. Our method performs joint estimation of functional labels and musical boundaries to alleviate the fragmentation issue.

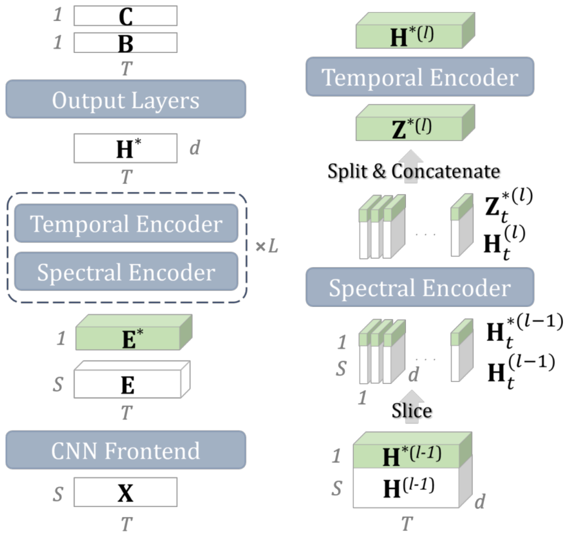
MHSA-based methods have recently been proposed for MSA owing to the excellent representation capability. The SpecTNT for MSA [27] uses Transformer encoders to capture the dependencies between the two axes of an input spectro-temporal representation. For training the SpecTNT with an increased amount of data, structure annotations from multiple datasets are mapped to the same semantic space with a 7-class taxonomy (‘intro’, ‘verse’, ‘chorus’, ‘bridge’, ‘inst’, ‘outro’, and ‘silence’). While the spectral components are often used for temporal modeling or collapsed before temporal modeling, this is the first attempt in the MSA task to retain the spectral dimension. In contrast, the convolution-augmented MHSA (CAMHSA) mechanism [28] captures temporal self-similarities on the self-attention maps derived from multiple types of acoustic features for capturing the repetitive nature of music. These network designs impose inductive biases that can enhance the representation learning.

Given the complementary aspects of these techniques, we integrate specialized MHSA mechanisms into the SpecTNT architecture for better modeling non-local features in spectral and temporal dimensions. Compared with the original CAMHSA [28], we retain the spectral dimension of the input data and aim to estimate the functional structure instead of the semiotic one, because the functional description conveys generic attributes of structural units that are comprehensible to the public. Compared with the original SpecTNT [27], we deal with spectro-temporal characteristics of music and processes input data at the track level rather than at the chunk (or segment) level, because the functional role of a structural unit might depend on the global organization of a musical piece.

3. PROPOSED METHOD

We tackle the functional MSA task with the 7-class taxonomy [27]. The estimation of the functional structure is formulated as a sequence labeling problem. Given a spectro-temporal representation, $\mathbf{X} \in \mathbb{R}^{T \times S}$, with S spectral components and T time steps, the goal of the estimation task is to output a sequence of categorical labels, $\mathbf{C} \in \mathbb{R}^{T}$, indicating the structural function of each time step $t \in T$. In practice, an extra binary sequence, $\mathbf{B} \in \{0, 1\}^{T}$, which specifies whether t is a boundary, is generated for smoothing the estimated labels within a segment surrounded by two boundaries. As depicted in Figure 2a, our model consists of three parts: a CNN-based frontend, L stacks of spectral and temporal encoders, and output layers in charge of the predictions of $\mathbf{B}$ and $\mathbf{C}$ respectively (in this paper, we use L=2 stacks for experiments). [1]

[1] The source code is available at https://github.com/Tsung-Ping/music-structure-analysis.



(a) (b)

Figure 2: (a) Model architecture. Extra spectral components $\mathbf{E}^*$ are stacked on the frontend output $\mathbf{E}$ before the encoder blocks. (b) Schematic diagram of the spectral and temporal encodings, where the initial input $\mathbf{H}^{(0)} = \mathbf{E}$, $\mathbf{H}^{*(0)} = \mathbf{E}^*$, and the final output $\mathbf{H}^{*(L)} = \mathbf{H}^*$. The spectral encoder represents the time slices ($[\mathbf{H}_t^*; \mathbf{H}_t]$) separately, and then the temporal encoder aggregates the extra components ($\mathbf{Z}_t^*$) and outputs a representation.

3.1 CNN Frontend

The frontend is composed of an initial stem with two 2-D convolutional layers followed by a residual block [29]. Let f_c denote a convolutional layer with d filters parameterized by θ. The outputs of the stem and the residual block, denoted by $\{\mathbf{X}', \mathbf{E}\} \in \mathbb{R}^{T \times S \times d}$, are computed as follows:

$$\mathbf{E} = f_c(f_c(\mathbf{X}', \theta_3), \theta_4) + \mathbf{X}', \quad (1)$$

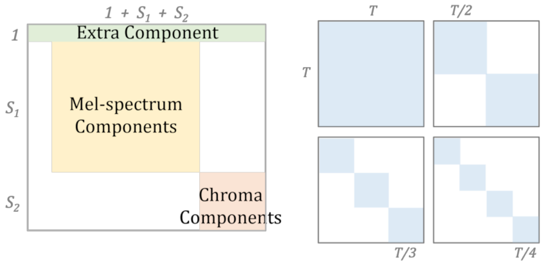
$$\mathbf{X}' = f_c(f_c(\mathbf{X}, \theta_1), \theta_2). \quad (2)$$

In effect, two frontends are employed to leverage different types of acoustic features. The two networks take as input the mel spectrogram ($\mathbf{X}_1 \in \mathbb{R}^{T \times S_1}$) and the chromagram ($\mathbf{X}_2 \in \mathbb{R}^{T \times S_2}$) respectively, and output $\mathbf{E}_1 \in \mathbb{R}^{T \times S_1 \times d}$ and $\mathbf{E}_2 \in \mathbb{R}^{T \times S_2 \times d}$. A unified representation is then obtained by concatenating the two outputs along the spectral dimension, i.e., $\mathbf{E} \in \mathbb{R}^{T \times (S_1 + S_2) \times d}$. We set $S_1 = 80$, $S_2 = 12$, and $d = 80$ for this work.

Note that adjacent pitch classes are irrelevant in the traditional sense of musical harmony, and we thus design the frontend for the chromagram carefully. Specifically, we concatenate $\mathbf{X}_2$ and the first 11 columns of $\mathbf{X}_2$ along the pitch-class axis, i.e., $\hat{\mathbf{X}}_2 = \mathrm{concat}(\mathbf{X}_2, \mathbf{X}_2[:, 1:11])$, and use convolutional layers with kernels that enclose the 12 pitch classes at once. This manipulation enables the CNN to capture key (or tonic)-independent patterns.

3.2 Spectral and Temporal Encoders

The spectral and temporal encoders are both built upon the Transformer encoder [10] that comprises stacks of MHSA blocks and feed-forward networks. Motivated by [28], we replace the standard MHSA of the temporal encoder with the CAMHSA mechanism. Moreover, we leverage relative position encodings [30] for enabling the model to process audio tracks of variable length.



(a) (b)

Figure 3: Attention masks for (a) the spectral MHSA and (b) the temporal MHSA. The rows (resp. columns) indicate *queries* (resp. *keys*) of the MHSA mechanism. Attention scores outside the colored regions will be filtered out.

As illustrated in Figure 2b, the two encoders in a stack are sequentially applied to the output of the previous stack. Let $\mathbf{E}_t \in \mathbb{R}^{S \times d}$ denote a time slice of $\mathbf{E}$ at t, and $\mathbf{E}_t^* \in \mathbb{R}^{1 \times d}$ a learnable vector that behaves as an extra spectral component. The spectral encoder (SE) jointly extracts latent features $\mathbf{Z}_t^* \in \mathbb{R}^{1 \times d}$ and $\mathbf{H}_t \in \mathbb{R}^{S \times d}$ from $\mathbf{E}_t^*$ and $\mathbf{E}_t$. Then, the temporal encoder (TE) exchanges information of $\mathbf{Z}_t^*$ across time as follows:

$$\mathbf{H}^* = [\mathbf{H}_1^{*(L)}; \ldots; \mathbf{H}_T^{*(L)}], \quad (3)$$

$$[\mathbf{H}_1^{*(l)}; \ldots; \mathbf{H}_T^{*(l)}] = \mathrm{TE}([\mathbf{Z}_1^{*(l)}; \ldots; \mathbf{Z}_T^{*(l)}]), \quad (4)$$

$$[\mathbf{Z}_t^{*(l)}; \mathbf{H}_t^{(l)}] = \mathrm{SE}([\mathbf{H}_t^{*(l-1)}; \mathbf{H}_t^{(l-1)}]), \quad (5)$$

where $\mathbf{H}^* \in \mathbb{R}^{T \times d}$ is the final output of the temporal encoder, $[\cdot;\cdot]$ denotes concatenation along the first dimension, l is the index of the stack, $\mathbf{H}_t^{*(0)} = \mathbf{E}_t^*$, and $\mathbf{H}_t^{(0)} = \mathbf{E}_t$. The extra spectral component $\mathbf{H}_t^*$ in each stack mimics the initial `CLS` token introduced in the BERT model [31] and is used to encapsulate spectral information at each time step. For a detailed description of the intertwined architecture, we refer the readers to [8].

To use the SE and TE for modeling spectral and temporal dependencies, we impose constraints on the attention maps of the MHSA block, as shown in Figure 3. For the SE, the attention between the two types of spectra (i.e., mel spectrum and the chroma features) and the attention on the extra component are masked out because such attentions would likely result in *diluted* representations [32]. While the between-type attentions are prohibited, their relations can be extracted via the extra component. For the TE, contextual information is aggregated simultaneously at four time scales (i.e., T, $T/2$, $T/3$, and $T/4$) in a *structure-aware* manner. Take the scale $T/2$ as an example, a time step $t < T/2$ can only attend the first half of the time axis. Considering that binary and ternary forms are common structures in Western music, such location-related information is expected to enhance the learning of music structure.

We leverage the multi-head nature of the MHSA block for simultaneous multi-scale attention.

3.3 Output Layers and Inference

The output layers consist of a three-layer fully connected neural network and take $\mathbf{H}^*$ as input to estimate the boundary likelihoods, $\mathbf{P}^{\mathrm{B}} \in [0,1]^T$, and the probability distributions over the 7 classes, $\mathbf{P}^{\mathrm{C}} \in [0,1]^{T\times 7}$, for all time steps:

$$\mathbf{P}^{\mathrm{B}} = \mathrm{sigmoid}(((\mathbf{H}^*\mathbf{W}_1^{\mathrm{B}})\mathbf{W}_2^{\mathrm{B}})\mathbf{W}_3^{\mathrm{B}}), \tag{6}$$

$$\mathbf{P}^{\mathrm{C}} = \mathrm{sigmoid}(((\mathbf{H}^*\mathbf{W}_1^{\mathrm{C}})\mathbf{W}_2^{\mathrm{C}})\mathbf{W}_3^{\mathrm{C}}), \tag{7}$$

where $\{\mathbf{W}_1^{\mathrm{B}}, \mathbf{W}_2^{\mathrm{B}}, \mathbf{W}_1^{\mathrm{C}}, \mathbf{W}_2^{\mathrm{C}}\} \in \mathbb{R}^{d\times d}$, $\mathbf{W}_3^{\mathrm{B}} \in \mathbb{R}^{d\times 1}$, and $\mathbf{W}_3^{\mathrm{C}} \in \mathbb{R}^{d\times 7}$ are learnable weight matrices. Note that we use the sigmoid function instead of the softmax activation in Eqn (7) for the function labeling is modeled as seven individual binary sequences.

To detect the boundaries $\mathbf{B}$ from $\mathbf{P}^{\mathrm{B}}$, we use a common peak-picking method [17] implemented in the librosa library [33]. To estimate the function labels $\mathbf{C}$, we give the segment between two adjacent boundaries (say t_1 and t_2) a label taking the largest average probability, i.e., $\mathbf{C}_t = \arg\max \sum \mathbf{P}_n^{\mathrm{C}} \;\; \forall\, t_1 \le t, n < t_2$.

3.4 Loss Function

Given the ground-truth boundaries and labels (represented by a sequence of one-hot vectors), $\mathbf{Y}^{\mathrm{B}} \in \{0,1\}^T$ and $\mathbf{Y}^{\mathrm{C}} \in \{0,1\}^{T\times 7}$ respectively, we compute the binary cross-entropy (BCE) losses for the model output to compute the overall loss ($\mathcal{L}$) as follows:

$$\mathcal{L} = \mathcal{L}^{\mathrm{B}} + \mathcal{L}^{\mathrm{C}}, \tag{8}$$

$$\mathcal{L}^{\mathrm{B}} = \mathrm{BCE}(\mathbf{Y}^{\mathrm{B}}, \mathbf{P}^{\mathrm{B}}), \tag{9}$$

$$\mathcal{L}^{\mathrm{C}} = \mathrm{BCE}(\mathbf{Y}^{\mathrm{C}}, \mathbf{P}^{\mathrm{C}}). \tag{10}$$

4. EXPERIMENTS

We conducted comparative experiments using the Beatles [34], RWC [35], and SALAMI [36] datasets. For the Beatles dataset, we used the refined Beatles-TUT annotations for 174 Beatles songs.[2] For the RWC dataset, we used the 100 songs from the Popular Music Database (denoted by RWC-POP). For the SALAMI dataset, we created a subset consisting of only popular music (SALAMI-POP), which amounted to 245 tracks. The maximum track length for each corpus was around 468 sec, 368 sec, and 438 sec. Following [27], we carried out cross-dataset evaluations for all the experiments. Each of the three corpora served as the test data in turn while the remainder was used for training. We augmented the training set via pitch shifting (within ±2 semitones) and pre-emphasis (with a coefficient of {0.7, 0.97}).

4.1 Statistics of the Function Labels

Structural annotations of the three corpora were converted to the 7-class label space with the mapping algorithm proposed in [27]. As illustrated in Figure 4, all the corpora were with concentrated distribution, where `Verse` and `Chorus` are the most common labels, since that verse-chorus form is widely used in popular music. It is also worth noting that `Inst` (i.e., 'instrumental') is extremely rare in the Beatles and RWC-POP datasets as a result of their annotation criteria and the mapping algorithm.

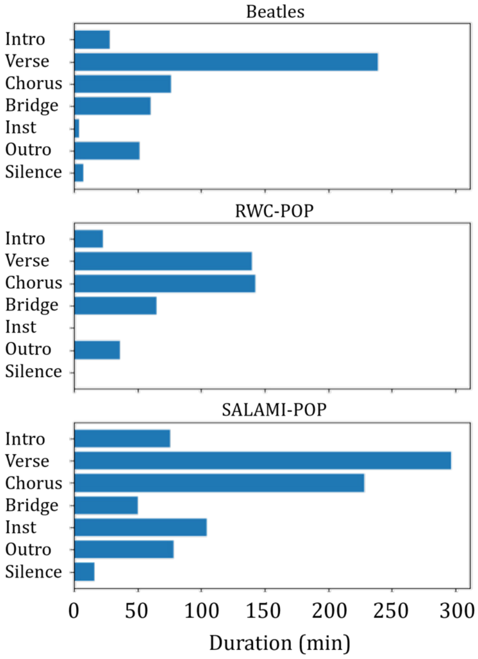


Figure 4: Statistics of the 7 function labels in each of the Beatles, RWC-POP, and SALAMI-POP corpora.

4.2 Input Representation

For each audio track, we computed the mel spectrogram with 80 mel bands and the chromagram with 12 chroma bins. The initial time resolution for both representations was 25 ms. We downsampled the two types of features by a factor of 20 (hence 1 frame = 0.5 sec) with the median filter so that the model could take as input a full-length track under memory constraints.

4.3 Evaluation Metrics

The performance on the MSA task was evaluated with the *mir_eval* library [37] in terms of segmentation and labeling. For segmentation, we computed the F1 score of the Hit Rate [38] with a time tolerance of ±0.5 sec and ±3 sec (denoted by HR.5F and HR3F respectively). For labeling, we computed the F1 score of the pairwise agreement [39] at the frame size of 0.1 sec (denoted by PWF).

In addition, the frame-wise labeling accuracy was measured in two ways. First, we converted the sequence of probabilities ($\mathbf{P}^{\mathrm{C}}$) into the labeling sequence ($\mathbf{C}$) either by taking the *argmax* function at each time step or by using the proposed *smoothing* strategy (Section 3.3). The derived labeling sequences were denoted by $\mathbf{C}_a$ and $\mathbf{C}_s$, respectively. Two types of labeling accuracy (ACC$_a$ and ACC$_s$) were then computed by comparing $\mathbf{C}_a$ and $\mathbf{C}_s$ with the

[2] `https://pythonhosted.org/msaf/datasets.html`.

Method	ACC_a/ ACC_s	HR.5F/ HR3F	PWF
	Beatles		
Proposed	**0.495/ 0.481**	**0.521/ 0.638**	0.571
ST-MHSA	0.410/ 0.386	0.480/ 0.610	0.576
TE-Only	0.455/ 0.451	0.484/ 0.610	0.547
SE-Only	0.355/ 0.330	0.448/ 0.600	**0.594**
	RWC-POP		
Proposed	**0.589/ 0.598**	**0.570/ 0.712**	**0.623**
ST-MHSA	0.425/ 0.426	0.498/ 0.637	0.537
TE-Only	0.528/ 0.531	0.504/ 0.662	0.578
SE-Only	0.428/ 0.430	0.472/ 0.644	0.562
	SALAMI-POP		
Proposed	**0.497/ 0.492**	**0.505/ 0.657**	**0.600**
ST-MHSA	0.411/ 0.401	0.435/ 0.559	0.561
TE-Only	0.422/ 0.411	0.452/ 0.582	0.575
SE-Only	0.425/ 0.390	0.418/ 0.492	0.552

Table 1: The result of the ablation study.

ground-truth labeling sequence, $\bar{\mathbf{C}} \in \{0, 1, \ldots, 6\}^T$:

$$\text{ACC}_\text{a} = \frac{1}{T}\sum_{t=1}^{T} \delta_{\bar{\mathbf{C}}_t, \mathbf{C}_{a,t}}, \tag{11}$$

$$\text{ACC}_\text{s} = \frac{1}{T}\sum_{t=1}^{T} \delta_{\bar{\mathbf{C}}_t, \mathbf{C}_{s,t}}, \tag{12}$$

where $\delta_{a,b}$ denotes the Kronecker delta function.

4.4 Baseline Methods

To validate the effectiveness of the spectral and temporal self-attentions, we conducted an ablation study with the following baseline models:

- **ST-MHSA**: Both the spectral and temporal encoders used the standard MHSA as in the SpecTNT [27].
- **TE-Only**: The spectral encoder was removed in a way similar to the CAMHSA work [28].
- **SE-Only**: The temporal encoder was removed from the proposed model. This was a localized prediction model taking only short-term context into account.

The ST-MHSA and TE-only are considered substitutes to the two previous works [27, 28], and we did not make a direct comparison to their results for a couple of reasons. First, the data used are different due to the difficulty of obtaining the exact audio signals: we used only the Beatles dataset while [27] used the Isophonics [34]; the subsets created from the SALAMI dataset are also different (245 tracks in our experiments and 274 tracks in [27]). Second, our model aims to predict semantic labels whereas the model of [28] outputs semiotic representations, and accordingly the data used are different.

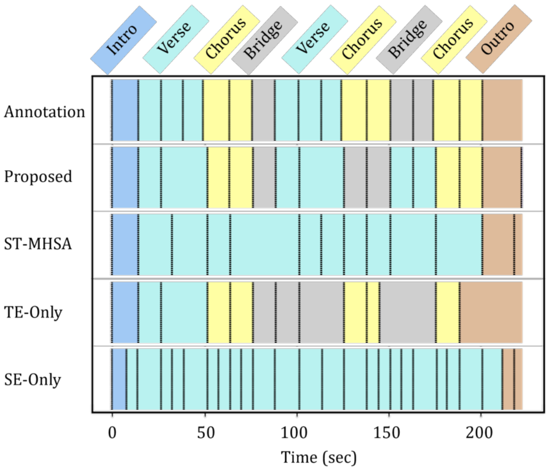


Figure 5: Structure analysis results of a song ("RM-P045") from the RWC-POP. The first row is the ground-truth annotation, and the other rows are the estimations by the proposed method and baseline models. The estimated boundaries are denoted by dashed lines.

To scrutinize our model design in relation to the performance, we also built variants of our model as follows:

- **w/ Pool**: The pooling operation was inserted into the CNN frontend for the mel spectrogram (Section 3.1). Precisely, we used a pooling layer for $\mathbf{X}'_1$ before the computation of Eqn (1). Following [12], we reduced the spectral dimension of the mel spectrogram using a max-pooling layer with a kernel size of 6 (while the temporal dimension was kept unchanged).
- **w/o S-Mask**: The spectral attention mask was not used (Section 3.2 and Figure 3a). This is equivalent to using a standard MHSA block for the SE.
- **w/o T-Mask**: The temporal attention mask was not used (Section 3.2 and Figure 3b). This is equivalent to using the CAMHSA mechanism for the TE.

5. RESULTS

We here report and discuss the results of the comparative and ablation experiments.

5.1 Comparison with Baseline Models

The results of the cross-dataset evaluations are summarized in Table 1. The proposed method outperformed the baseline methods on the three corpora in most metrics (with a comparable PWF score to the ST-MHSA and the TE-Only on the Beatles). In comparison with the ST-MHSA, the performance gain of the proposed method was mainly attributed to the tailored MHSA blocks for spectral and temporal modelings, validating the importance of the MHSA adaptation to the task. Given that the TE-Only obtained better ACC scores than the SE-Only, the temporal self-attention was considered to have a greater impact on identifying structural functions than its spectral counterpart.

Method	ACC_a/ ACC_s	HR.5F/ HR3F	PWF
	Beatles		
Proposed	0.495/ 0.481	0.521/ 0.638	0.571
w/ Pool	0.387/ 0.370	0.446/ 0.566	0.536
w/o S-Mask	0.498/ 0.497	0.520/ **0.654**	0.582
w/o T-Mask	**0.538/ 0.531**	**0.528/** 0.643	**0.614**
	RWC-POP		
Proposed	**0.589/ 0.598**	0.570/ 0.712	**0.623**
w/ Pool	0.528/ 0.489	0.395/ 0.550	0.546
w/o S-Mask	0.571/ 0.577	**0.571/ 0.715**	0.621
w/o T-Mask	0.576/ 0.567	0.549/ 0.686	0.607
	SALAMI-POP		
Proposed	**0.497/ 0.492**	**0.505/ 0.657**	**0.600**
w/ Pool	0.495/ 0.454	0.418/ 0.522	0.552
w/o S-Mask	0.487/ 0.490	0.480/ 0.628	0.581
w/o T-Mask	0.471/ 0.480	0.485/ 0.635	0.586

Table 2: Evaluations of the model design choices.

Nonetheless, the differences between the SE-Only and TE-Only were not clear in terms of the HR and the PWF scores, implying that spectral and temporal self-attentions both can contribute to the tasks. In addition, we found that our inference strategy smoothed the structural labeling results while having a minor effect on the ACC score.

Figure 5 portrays the structural labeling results for one song from the RWC-POP corpus. Regarding the segmentation results, all the models were able to detect the transitions between different sections, but the finer structure (i.e., repetitions or variants within a coarse section) was sometimes overlooked. In particular, the SE-Only over-segmented the musical track due to the limited contextual information. Regarding the labeling results, the proposed method and the TE-Only were capable of correctly estimating the five function labels in the track, whereas the ST-MHSA and SE-Only failed to identify all the chorus and bridge sections, possibly owing to the insufficient capability of temporal modeling.

5.2 Evaluation of Design Choices

Experiments results regarding the model design are listed in Table 2. As we expected, the severe performance degradation was caused by the pooling operation (w/ Pool) on the three corpora. Spectral components are not pixels that are highly correlated in local regions, and therefore naive local pooling could be detrimental to spectral features. As for the attention masking (w/o {S, T}-MASK), the results suggested that the imposed constraints can have a positive impact on the performance. On the SALAMI-POP, which is the most challenging one among the three corpora, the MHSA mechanism without any constraints resulted in a clear performance drop. In particular, we found that unconstrained spectral components tended to give great attention to the extra component ($\mathbf{H}_t^*$) rather than themselves. A similar effect was also reported by previous research in the field of natural language processing [40,41]. This kind of concentrated attention to a special (or artificial) component that has distinct semantic meanings could downplay the representation capability of the MHSA.

5.3 Evaluation of Spectro-Temporal Self-Attentions

The attention maps implicitly computed with the MHSA mechanism often disclose illuminating relationships between input elements [42–45]. The spectral and temporal self-attentions of our model also exhibited such an effect. As depicted in Figure 1c, the leftmost temporal self-attention map highlighted a potential musical event at around 66 sec, which could be associated with the variant repeat of the first 10 sec of this chorus section (as can be seen in Figure 1a, two triangular patterns span from 56 to 66 sec and from 66 to 76 sec, respectively). This result echos the observation that self-attention maps can represent music structure [28]. In contrast, the spectral self-attention, as illustrated in Figure 1b, uncovered the tonal relationships between the 12 pitch classes with an emphasis on the notes comprising the tonic chord (assume in the key of C major). Particularly, pitch class E gained persistent attention over this section even though it had a low energy level for most of the time. Through alternate self-attention across the spectral and temporal dimensions, the contextual information of individual aspects can be mingled effectively and provide insights into music structure.

6. CONCLUSION

We have presented a deep learning model for music structure analysis, especially from the perspective of the functional structure representation. The core idea of this study is to learn non-local spectral and temporal dependencies inherent in music with clear distinction. For this purpose, we adapted the multi-head self-attention mechanism for each aspect and leveraged two types of the Transformer encoder to unravel the spectro-temporal relationships. Compared with the ablated variants of the Transformer encoder, the proposed model with the specialized self-attention mechanisms worked better on three datasets in music segmentation and structure labeling. The learned self-attention maps unveiled that the correlations between separated spectral or temporal components can be effective clues for modeling music structure.

In spite of these encouraging results, we acknowledge the computational limitation of our approach. Apart from an M-head temporal self-attention having the memory footprint of $M \times T \times T$, an N-head spectral self-attention involves intermediate attention maps with $T \times N \times S \times S$ space complexity. Given that our method aims to process full-length audio data (hence larger T) and leverage multiple types of acoustic features (hence greater S), memory-efficient self-attention mechanisms are critical to this kind of dual-axis modeling. Time and frequency are intricately interwoven to form the musical fabric, and each individual aspect is worth considerable attention.

7. ACKNOWLEDGMENT

This work was partially supported by JST FOREST No. JPMJFR2270 and JSPS KAKENHI Nos. JP24H00742, JP24H00748, and JP24KJ1379.

8. REFERENCES

[1] O. Nieto, G. J. Mysore, C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 246–263, 2020.

[2] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, "Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 235–240.

[3] T. Grill and J. Schlüter, "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1296–1300.

[4] A. Maezawa, "Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 206–210.

[5] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, "Learning multi-level representations for hierarchical music structure analysis." in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 591–597.

[6] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW)*, 2016, p. 125.

[8] W. T. Lu, J. Wang, M. Won, K. Choi, and X. Song, "SpecTNT: a time-frequency Transformer for music audio," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 396–403.

[9] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in Transformer," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 15 908–15 919.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[11] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 547–553.

[12] K. Ullrich, J. Schlüter, and T. 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, pp. 417–422.

[13] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 531–537.

[14] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 346–350.

[15] C. Hernandez-Olivan, J. R. Beltrán, and D. Diaz-Guerra, "Music boundary detection using convolutional neural networks: A comparative analysis of combined input features," *International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI)*, vol. 7, no. 2, p. 78, 2021.

[16] G. Peeters, "Self-similarity-based and novelty-based loss for music structure analysis," in *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 749–756.

[17] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 49–54.

[18] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[19] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 405–410.

[20] C. Wang and G. J. Mysore, "Structural segmentation with the Variable Markov Oracle and boundary adjustment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 291–295.

[21] T. Cheng, J. B. L. Smith, and M. Goto, “Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 106–110.

[22] C. J. Tralie and B. McFee, “Enhanced hierarchical music structure annotations via feature level similarity fusion,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 201–205.

[23] A. Marmoret, J. E. Cohen, and F. Bimbot, “Barwise music structure analysis with the correlation block-matching segmentation algorithm,” *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 6, no. 1, pp. 167–185, 2023.

[24] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, “Self-supervised learning of multi-level audio representations for music segmentation,” *IEEE ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 32, pp. 2141–2152, 2024.

[25] G. Shibata, R. Nishikimi, and K. Yoshii, “Music structure analysis based on an LSTM-HSMM hybrid model,” in *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 23–29.

[26] J. Wang, J. B. L. Smith, J. Chen, X. Song, and Y. Wang, “Supervised chorus detection for popular music using convolutional neural network and multi-task learning,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 566–570.

[27] J. Wang, Y. Hung, and J. B. L. Smith, “To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 416–420.

[28] T. Chen, L. Su, and K. Yoshii, “Learning multifaceted self-similarity for musical structure analysis,” in *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 165–172.

[29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 464–468.

[31] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional Transformers for language understanding,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[32] S. Mehri and M. Eric, “Example-driven intent prediction with observers,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, 2021, pp. 2979–2992.

[33] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in Python,” in *Proceedings of the 14th Python in Science Conference (SciPy)*, 2015, pp. 18–24.

[34] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, “OMRAS2 metadata project 2009,” in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR) - Late-Breaking Session*, 2009.

[35] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical and jazz music databases,” in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, 2002.

[36] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. 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, pp. 555–560.

[37] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 367–372.

[38] D. Turnbull, G. R. G. Lanckriet, E. Pampalk, and M. Goto, “A supervised approach for detecting boundaries in music using difference features and boosting,” in *Proceedings of the 8th International Conference on Music Information (ISMIR)*, 2007.

[39] M. Levy and M. B. Sandler, “Structural segmentation of musical audio by constrained clustering,” *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[40] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does BERT look at? An analysis of BERT’s attention,” in *Proceedings of the 2019 ACL Workshop BlackboxNLP*, 2019, pp. 276–286.

[41] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, “Revealing the dark secrets of BERT,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4364–4373.

[42] C. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, “Music Transformer: Generating music with long-term structure,” in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[43] J. Vig and Y. Belinkov, “Analyzing the structure of attention in a Transformer language model,” in *Proceedings of the ACL Workshop BlackboxNLP*, 2019, pp. 63–76.

[44] T. Chen and L. Su, “Attend to chords: Improving harmonic analysis of symbolic music using Transformer-based models,” *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 4, no. 1, pp. 1–13, 2021.

[45] K. Shim, J. Choi, and W. Sung, “Understanding the role of self attention for efficient speech recognition,” in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.