TISMIR

Chen, T.-P., and Su, L. (2021). Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 1–13. DOI: https://doi.org/10.5334/tismir.65

RESEARCH

Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models

Tsung-Ping Chen and Li Su

Automatic chord recognition (ACR) has long been a topic of interest in the field of Music Information Retrieval (MIR), due to not only its commercial applications, but also its support for advanced music analysis. While a lot of ACR-related work deals with audio data, ACR from symbolic music has received less attention. In addition, conventional ACR systems specify chords in a key-dependent way (usually with the root note and the chord quality) and hence are unable to reveal the high-level patterns and harmonic structures. These issues hinder the developments of music analysis and music generation via ACR systems. With the success of deep learning, it is viable to build a symbolic ACR system using a more comprehensive chord vocabulary such as functional harmony. Recently, two advanced models, namely the Bi-directional Transformer for Chord Recognition (BTC) and the Harmony Transformer (HT), introduced for the first time the multi-head attention mechanism to ACR, showing the great capability of the attention mechanism to improve the performance of ACR. In this paper, we systematically study the performance of the BTC and the HT in terms of symbolic ACR, and propose an improved model. Experiments on conventional ACR and advanced functional harmony recognition indicate that the HT has the potential to surpass the BTC, especially in terms of chord segmentation quality. Also the overall performance of the HT is further improved by enhancing the learning of local context and positional information.

Keywords: automatic chord recognition; functional harmony recognition; symbolic music; Transformer; multi-head attention; chord segmentation

1. Introduction

1.1 Automatic Chord Recognition

Chord recognition is a process to identify the harmonic entity of each musical segment, usually by giving a chord name to the segment in question. This problem is not as simple as it may seem, for it concerns several aspects of musical harmony, and the answer to the problem may not be unique. For instance, a C major sixth chord, `C6`, in popular music may be termed as an A minor seventh chord in first inversion, `Am7/C`, in classical music. In other words, the chord vocabulary differs according to musical style and context. Moreover, the boundary of each segment which deserves to be recognized as a single chord is not explicitly defined by the music itself. Therefore, it is usually difficult to partition music into harmonically meaningful segments. A comprehensive chord recognition approach should specify the information of when to identify a chord and how to build the chord. Owing to its complexity, ACR is still challenging even nowadays.

Institute of Information Science, Academia Sinica, TW

Corresponding author: Tsung-Ping Chen (tearfulcanon@iis.sinica.edu.tw)

1.2 ACR in Audio Domain

During the past decades, researchers have studied ACR, particularly for audio data, from different aspects, such as chord segmentation (Yoshioka et al., 2004; Harte et al., 2006; Degani et al., 2015), beat and key (Zenz and Rauber, 2007), bass line (Yang et al., 2016), root notes (Yang et al., 2016), and meter (Degani et al., 2017). Although there are various approaches to ACR, much of the work targets calculation of effective chroma features for comparison with a set of chord templates (Fujishima, 1999; Lee, 2006; Stark and Plumbley, 2009; Oudre et al., 2011). Based on the chroma features, stochastic methods such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) are frequently employed to predict chord progressions (Sheh and Ellis, 2003; Cho and Bello, 2009; Ueda et al., 2010; Deng and Kwok, 2016; Korzeniowski and Widmer, 2016). This combination of audio feature extraction and chord sequence prediction can be analogous to the integration of acoustic modeling and language modeling in the field of speech recognition (Li and Wu, 2015), and has become a typical framework for ACR from audio.

Along with the rise of deep learning, ACR has experienced a paradigm shift from template-based algorithms to data-driven approaches. Works on ACR have begun to explore the

capability of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for extracting non-handcrafted features and for modeling chord sequences (Humphrey and Bello, 2012; Boulanger-Lewandowski et al., 2013; McFee and Bello, 2017; Hori et al., 2017). Besides, various chord vocabularies ranging from basic triads (Zhou and Lerch, 2015; Korzeniowski and Widmer, 2016) to more complex chords (Humphrey and Bello, 2015; Deng and Kwok, 2017; Jiang et al., 2019) have been taken into consideration in ACR systems. However, most of the audio datasets commonly used in ACR work provide only chord labels without the audio data because of copyright restrictions, preventing evaluation on the same benchmarks. Readers are recommended to refer to Pauwels et al. (2019) for a thorough review of ACR from audio during the past two decades.

1.3 ACR in the Symbolic Domain

In comparison to ACR from audio data, ACR from symbolic music has received less attention (Scholz and Ramalho, 2008; Rhodes et al., 2009; Rocher et al., 2009; Masada and Bunescu, 2019). One reason is that the public are more likely to access audio content rather than symbolic music. Another reason is that high-quality symbolic music data and the corresponding annotations are relatively scarce. The difference between audio ACR and symbolic ACR mainly lies in the processing of the music data. Audio music data contain expressive information (e.g., timbre) and are often represented as spectrograms or chromagrams using the Short-Time Fourier Transform (STFT) or the constant-Q transform (CQT). On the other hand, symbolic music data comprise abstract concepts of music (e.g., the pitch of each single note), and hence can be represented as sets of note-related features (Masada and Bunescu, 2017) or piano rolls (Chen and Su, 2018). In spite of such a difference, symbolic ACR and audio ACR are capable of achieving competitive performances with nearly identical neural network architectures, since the two types of data can be represented and structured in a similar way (Chen and Su, 2019). It has to be mentioned that symbolic ACR using deep learning approaches is still in its preliminary stage, and previous research employed different datasets (usually comprising limited amounts of data) for evaluation. Therefore, more systematic studies are needed to explore the capability of deep neural networks and to set benchmarks for this field.

The development of symbolic ACR is valuable from four points of view. First, while the majority of research by musicologists and music theorists focuses on symbolic music representations, relatively few symbolic ACR and related MIR tools are available. Second, symbolic music data are invariant to some aspects of musical interpretation and thus can reduce the bias resulted from the performing and the recording factors. Third, the analysis of symbolic music data can be easily related to the abstract aspect of music by which musicological insights are derived. Finally, the growing interest in music generation will yield a lot of symbolic music data (Dong and Yang, 2018; Donahue et al., 2019; Lim et al., 2020); hence an increase in demand for symbolic ACR would likely occur before long.

1.4 Functional Harmony Recognition

Functional harmony is an advanced way of describing chords, typically by performing Roman numeral (RN) analysis. Traditionally, RN analysis includes the specification of key (or tonic), modulation, chord quality, chord inversion, and chord alteration. Instead of merely giving a chord name to each harmonic entity, as is usually done in ACR work, RN analysis represents chords as RNs which indicate the root notes of the chords in relation to the tonic, and specify the harmonic functions (hence the term *functional harmony*). As a result, the RN representation is key-invariant, and provides more informative analysis of musical harmony. According to the aforementioned facts, the recognition of functional harmony involves more than identifying chords, and is more demanding than the conventional ACR.

While some research has contributed to functional harmony recognition (Tsui and MacLean, 2002; Raphael and Stoddard, 2004; Illescas et al., 2007; Passos et al., 2009; de Haas et al., 2011), very few of them employed deep learning methods, partly due to the limited amounts of training data. Fortunately, several symbolic corpora containing RN annotations have been published in the past few years, e.g., the TAVERN dataset (Devaney et al., 2015), the ABC dataset (Neuwirth et al., 2018), the BPS-FH dataset (Chen and Su, 2018), and the Bach Preludes (Tymoczko et al., 2019). In addition, researchers have begun to address the functional harmony recognition task using advanced deep learning techniques (Chen and Su, 2019; Micchi et al., 2020). Considering that functional harmony is a fundamental way to find common patterns in chord progressions and to uncover the tonal structure of a musical piece, research on harmony recognition will benefit both the symbolic MIR and the musicology communities (Gotham and Ireland, 2019).

1.5 Attend to the Chords

As in language, chord progressions and musical harmony are highly context-dependent. Therefore, it is important to consider the relations between harmonic entities when recognizing chords and their functions. Although RNNs are representative models for capturing contextual information, the attention mechanism has shown its potential for sequence modeling (Bahdanau et al., 2015; Hermann et al., 2015; Parikh et al., 2016). The first fully attention-based model, namely the Transformer (Vaswani et al., 2017), was accordingly proposed to compete with the previously existing models of sequence learning. Its variant, BERT (Bi-directional Encoder Representations from Transformers) (Devlin et al., 2019), was devised to function as a pre-trained model for learning word representations. These attention-based models made a lot of breakthroughs in language-related tasks, and stood as landmarks in natural language processing (NLP). Due to its promising performance, the attention mechanism has been applied to many other tasks beyond the realm of NLP, such as image and music generation (Parmar et al., 2018; Huang et al., 2019).

Recently, two Transformer-based models, the bi-directional Transformer for chord recognition (BTC)

(Park et al., 2019) and the Harmony Transformer (HT) (Chen and Su, 2019), were proposed for the first time to tackle the ACR task. The BTC utilized a self-attention mechanism to capture the long-term dependency in musical sequences, and showed its ability to segment chord sequences; the HT estimated chord transitions (or chord boundaries), and then recognized chords via attending to the segmentation-informed sequence.[1] Although the two models were built upon the Transformer, they differed from each other in two aspects. First, the BTC utilized the encoder part of the Transformer, while the HT employed the entire Transformer architecture. Second, the BTC was trained on audio datasets, while the HT was applied to both audio and symbolic music data. In spite of these differences, the two models have demonstrated the effectiveness of the attention mechanism on modeling chord progressions, and outperformed other promising models in previous research (Korzeniowski and Widmer, 2016, 2017; McFee and Bello, 2017). Since the attention mechanism can access all positions of a sequence at a time, the Transformer-based models are theoretically more capable of capturing thorough knowledge of the sequence than RNNs, and therefore may alleviate the issue of temporal fragmentation when modeling chord sequences at the time-frame level (Korzeniowski and Widmer, 2017).

In this work, we tackle symbolic ACR using Transformer-based models, and propose improvements on the models. With evaluations on the conventional chord recognition and the functional harmony recognition tasks, we show that the HT is more promising than the BTC in terms of recognition accuracy and segmentation quality; it is also validated that the proposed improvements advance the overall performance of ACR. The major contribution of this paper is to propose an improved Transformer-based network for symbolic ACR, specifically on the challenging functional harmony recognition task, via a systematic investigation on the symbolic ACR research and Transformer-based ACR methods. The remainder of the paper is arranged as follows: we first introduce the Transformer, examine the architectures of the HT and the BTC, and propose methods to advance the performance of chord recognition (Section 2). Then we conduct experiments to evaluate the models and the proposed methods using two symbolic datasets (Section 3). The future work is then discussed (Section 4). Finally, the concluding remarks are presented (Section 5).

2. Transformer for Chord Recognition

2.1 Building Blocks of Transformer

The Transformer comprises two major computational blocks: the multi-head attention (MHA) and the feed-forward network (FFN). For the MHA, the concept of key-value memory is adopted, in which the *keys* are used to address relevant memories with respect to a *query*, and the corresponding *values* are subsequently returned (Miller et al., 2016). For example, given a query to search for a document in a database, the search engine will map the query against a set of keys (e.g., title, description, etc.) associated with candidate documents in the database, then present the best matched documents (values). In the case of sequence-to-sequence learning, the queries stand for the target sequences, while the keys and the values both represent the source sequences. Specifically, given queries (**Q**) and a set of key-value pairs (**K, V**), an attention function can be generalized to compute a weighted sum of the values (i.e., a context vector) for each query according to the relations between the query and the corresponding keys. By separately dividing **Q, K, V** into partitions, the MHA applies multiple independent attention functions (hence *multi-head*) to the partitions and extracts various context vectors ($\mathbf{C}_i$):

$$\begin{aligned}\mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathrm{Concat}(\mathbf{C}_1,\cdots,\mathbf{C}_h)\mathbf{W}^{\mathrm{C}},\\ \mathbf{C}_i &= \mathrm{Attention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i)\\ &= \mathrm{softmax}\left(\frac{(\mathbf{Q}_i\mathbf{W}_i^{\mathrm{Q}})(\mathbf{K}_i\mathbf{W}_i^{\mathrm{K}})^{\top}}{\sqrt{d}}\right)(\mathbf{V}_i\mathbf{W}_i^{\mathrm{V}}),\end{aligned} \tag{1}$$

where $\mathbf{W}^{\mathrm{C}}$, $\mathbf{W}_i^{\mathrm{Q}}$, $\mathbf{W}_i^{\mathrm{K}}$, and $\mathbf{W}_i^{\mathrm{V}}$ are learnable parameter matrices, h is the number of heads, i indicates the ith partition, and d is the feature size of $\mathbf{K}_i$.

According to the inputs of the function, we differentiate two types of MHA:

- **Inter-MHA: Q** and **K** represent different sequences, e.g., the target sentences (**Q**) and the source sentences (**K**).
- **Intra-MHA: Q** and **K** represent the same sequences; also known as *self-attention*.

Regarding the valid positions (or time steps) of a sequence to which the attention function is applied, MHA can be classified into another two categories:

- **Bi-directional MHA:** all positions are valid.
- **Uni-directional MHA:** each position can only attend to those positions either preceding itself (backward) or succeeding itself (forward).

For the Transformer, as shown in **Figure 1a**, the MHA units in the encoder layers and the first MHA unit in each decoder layer are intra-MHAs, while the second MHA unit in each decoder layer is an inter-MHA. It is worth noting that the intra-MHAs in the decoder are uni-directional (backward) due to the autoregressive decoding approach of the Transformer.

On the other hand, the FFN unit is composed of 2 fully-connected layers by which the input is projected onto a higher dimensional space followed by a rectified linear unit (ReLU), and then projected back to its original dimensions:

$$\mathrm{FFN}(\cdot) = \mathrm{ReLU}(\cdot\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{2}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are parameter matrices, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are learnable bias vectors. Alternatively, the fully-connected layers in the FFN unit can be replaced with 1-dimensional convolutional neural networks:

$$\mathrm{FFN}'(\cdot) = \mathrm{ReLU}(\cdot * \mathbf{W}_1' + \mathbf{b}_1') * \mathbf{W}_2' + \mathbf{b}_2', \tag{3}$$

where * denotes the convolution operation. We refer to the two variants of FFN as fully-connected FFN and convolutional FFN. The sizes of all the parameter matrices are shown in **Table 1**.

In practice, the MHA and the FFN both employ residual connections (He et al., 2016) followed by layer normalization (Ba et al., 2016). Moreover, since the two computational units process all positions of a sequence in a parallel manner without regard to the sequential order, Transformer-based models usually incorporate explicit position information, such as absolute positional encodings (Vaswani et al., 2017) and relative positional encodings (Shaw et al., 2018; Dai et al., 2019; Huang et al., 2019).

2.2 BTC versus HT

The BTC is built upon the Transformer encoder in a fashion similar to BERT, aiming to learn the *bi-directional* representations of the inputs. Specifically, the BTC learns the representations by employing two *uni-directional* intra-MHAs, one forward and the other backward, both followed by a convolutional FFN, as depicted in **Figure 1b**. This bi-directional approach makes a distinction from BERT in which the bi-directional intra-MHA is used. Although the technique of combining forward and backward intra-MHAs has been practiced in the fields of NLP and computer vision (Shen et al., 2018; Hossain et al., 2019), its effect on music data is still unexplored.

In contrast to the BTC, the HT retains the encoder-decoder architecture for the sake of integrating the chord change prediction into the chord recognition process. As shown in **Figure 1c**, the encoder and the decoder are the same as that of the Transformer, except that the intra-MHA of the decoder is bi-directional rather than uni-directional (backward). This difference results from the fact that the HT adopts a non-autoregressive framework (Gu et al., 2018), and hence can dismiss the backward constraint. Another difference is that the encoder of the HT has an additional output for the chord change prediction.

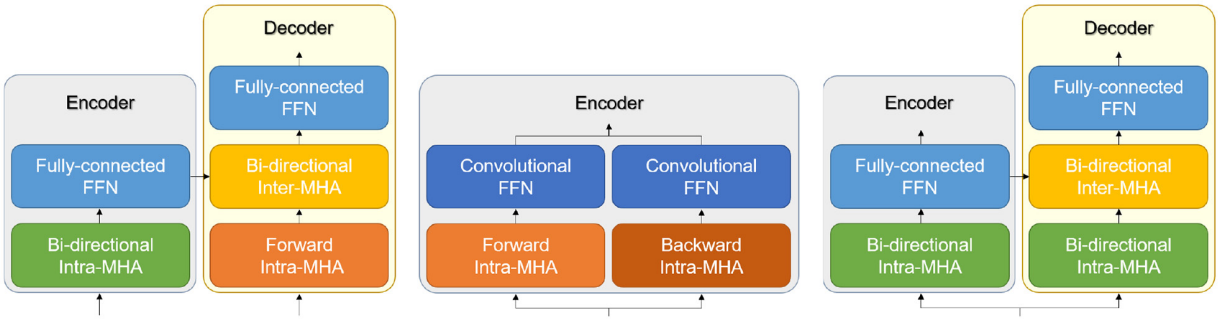
Comparatively speaking, the HT is more complex than the BTC in terms of model architecture and training techniques. In the aspect of model architecture, the HT has one additional inter-MHA in order to connect the decoder with the encoder. As a result, the HT has slightly more parameters than the BTC (around $4h^2d^2$). In practice, however, the BTC consists of more repetitive layers than the HT (8 and 2, respectively), resulting in higher model capacity. In the aspect of training techniques, the HT includes a computational block called *regionalization* (not shown in **Figure 1c**) to pass the chord change prediction to the decoder; also, softmax-normalized layer weights are employed to compute the weighted sum of the outputs from all repeated layers.

2.3 Improving the Transformer-Based Models

Both the BTC and the HT process the input sequences at the frame level. That is, the intra-MHAs at the bottom of the models represent each frame with respect to the relations between the *individual* frames. However, the frame sizes (which may be on the scale of 100 milliseconds for audio

Table 1: Number of parameters in the MHA and the FFN blocks; h, d, and n stand for the number of heads, the feature size of the partitioned keys ($\mathbf{K}_i$), and the kernel size of the convolution, respectively. We set $h = 4$, $d = 32$, and $n = 3$ for the experiments.

Computational Block	Parameter	Size	Total
Multi-head Attention	$\mathbf{W}^{\mathrm{C}}$	$hd \times hd$	$4h^2d^2$
	$\mathbf{W}_i^{\mathrm{Q}}, \mathbf{W}_i^{\mathrm{K}}, \mathbf{W}_i^{\mathrm{V}}$	$hd \times d$	
Fully-connected FFN	$\mathbf{W}_1$	$hd \times 4hd$	$8h^2d^2$
	$\mathbf{W}_2$	$4hd \times hd$	
Convolutional FFN	$\mathbf{W}_1', \mathbf{W}_2'$	$n \times hd \times hd$	$2nh^2d^2$



(a) Transformer **(b)** BTC **(c)** HT

Figure 1: Basic building blocks of the Transformer, the bi-directional Transformer for chord recognition (BTC), and the Harmony Transformer (HT) models, using on multi-head attention (MHA) and feed-forward networks (FFN). Note that both the encoder and the decoder have repetitive layers which are not shown in the figure.

and of a 16th note for symbolic music) are somewhat small for encoding harmonic content. Moreover, the MHAs access all the time steps of a sequence in parallel at the expense of knowing the sequential order. Although the BTC and the HT employed the *absolute* positional encodings of the Transformer in compensation, it was argued that the *relative* differences in position matter more for music (Huang et al., 2019).

We propose to improve the Transformer-based ACR models according to the two aforementioned aspects. First, we introduce intra-block self-attention (Shen et al., 2018), or intra-block intra-MHA, to the input of the models for learning localized harmonic features. Concretely, the intra-block intra-MHA unit splits a sequence into B blocks of equal length m and captures the local dependency within each block with the bidirectional intra-MHA:

$$\begin{aligned}&\text{Intra-block Intra-MHA}(\mathbf{Q},\mathbf{K},\mathbf{V})\\&=\begin{pmatrix}\text{Intra-MHA}(\mathbf{Q}^{(1)},\mathbf{K}^{(1)},\mathbf{V}^{(1)})\\\vdots\\\text{Intra-MHA}(\mathbf{Q}^{(B)},\mathbf{K}^{(B)},\mathbf{V}^{(B)})\end{pmatrix},\end{aligned}\tag{4}$$

where $\mathbf{X}^{(b)}$ indicates the bth block of $\mathbf{X}$, and $(\cdot)$ denotes a concatenation along the time dimension. In other words, the intra-MHA is applied to the B blocks individually for modeling the local context inside each of them. Second, we employ relative positional encodings (Dai et al., 2019) and positional attention (Gu et al., 2018) to enhance the model's knowledge of sequential order. In this way, the relative positional encodings enable the MHA to consider the pairwise relationships between its input elements, while the positional attention incorporates positional information directly into the attention process.

We apply the aforementioned techniques to the HT, and the architecture of the derivative model (denoted as HT*) is illustrated in **Figure 2**. The intra-block intra-MHAs (followed by convolutional FFNs) are added to the bottoms of the encoder and the decoder; the relative position encodings are introduced to all the MHAs; and the positional attention is employed in the decoder according to its original setting. In addition, the fully-connected FFNs of the HT are replaced by the convolutional FFNs in order to capture the adjacent information of their inputs before outputting to the next repetitions. From the perspective of the network topology, the combination of the intra-block intra-MHA and the bi-directional intra-MHA functions in a way similar to a convolutional recurrent neural network (CRNN) which captures local features with convolutions first and models long-term structure with recurrences thereafter (McFee and Bello, 2017; Micchi et al., 2020). We use $m = 4$ for the intra-block intra-MHA, and set the maximum relative position to the input sequence length. In the following section, we experimentally compare the BTC and the HT, and validate the improvement in the HT*.

3. Experiments

3.1 Testing Corpora

To train the models for the chord recognition and the functional harmony recognition tasks, two corpora are used: the BPS-FH dataset and the Bach Preludes, where symbolic music and human-annotated RN labels are provided. The former includes complete first movements of the Beethoven Piano Sonatas (32 movements in total), and the latter consists of 24 preludes from the first book of Bach's Well Tempered Clavier. We use the BPS-FH dataset because it was used to evaluate the HT. In addition, we include the Bach Preludes as it has similar properties to the BPS-FH dataset (both comprise piano solos). We leave other datasets (e.g., the ABC dataset and the TAVERN dataset) for future work.

The analytic information of the Bach Preludes is transcribed into the tabular format of the BPS-FH for unifying the notation system of the two corpora. We additionally derive chord symbols from the RN annotations for the chord recognition task. In sum, there are 11478 labels in the BPS-FH, and 2615 labels in the Bach Preludes. All analyzed chords in the two corpora are categorized

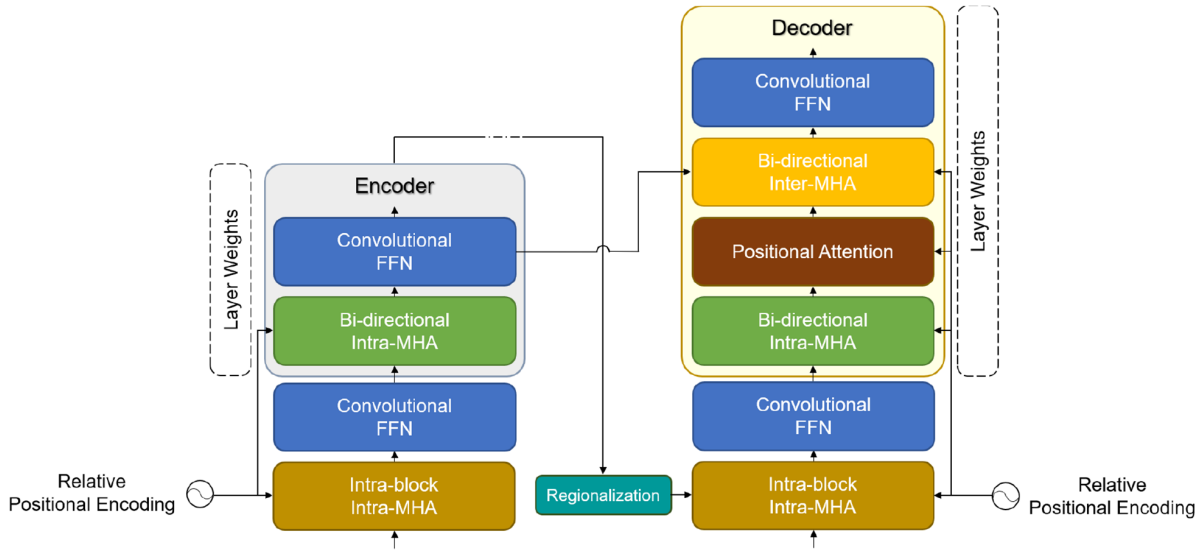


Figure 2: Improved Harmony Transformer (HT*).

into 10 classes, as shown in **Table 2**. **Figure 3** depicts the statistics of the annotated chords of the two corpora in terms of chord quality and degree. The statistics of chord quality show that major, minor and dominant seventh take the majority in both corpora, followed by minor seventh, diminished, and diminished seventh. And the degree distributions also manifest a similarity between the two corpora, in the sense that the top six chord degrees (i.e., `1,5,2,4,7,6`) are almost the same.

Table 2: Annotated chord qualities and the mapping to the major-minor vocabulary.

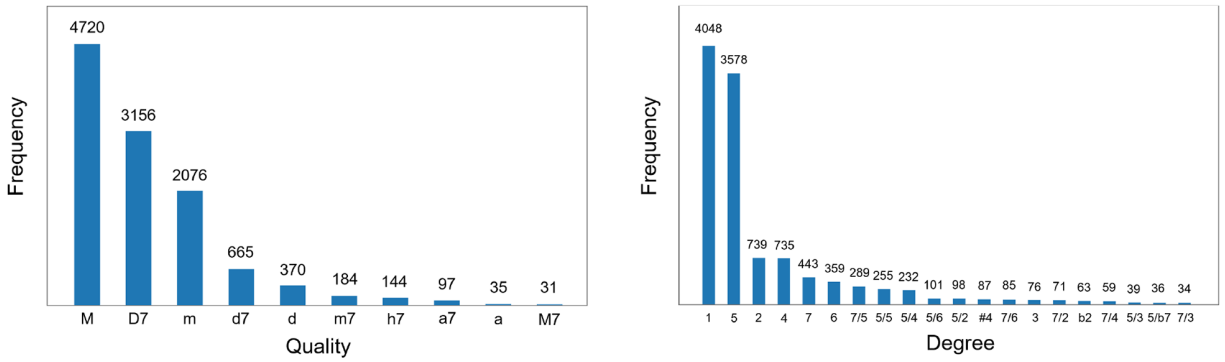
Quality	Major-Minor Mapping
Major (M)	M
Minor (m)	m
Augmented (a)	others
Diminished (d)	others
Major Seventh (M7)	M
Minor Seventh (m7)	m
Dominant Seventh (D7)	M
Diminished Seventh (d7)	others
Half-diminished Seventh (h7)	others
Augmented Sixth (a6)	M

3.2 Data Representation

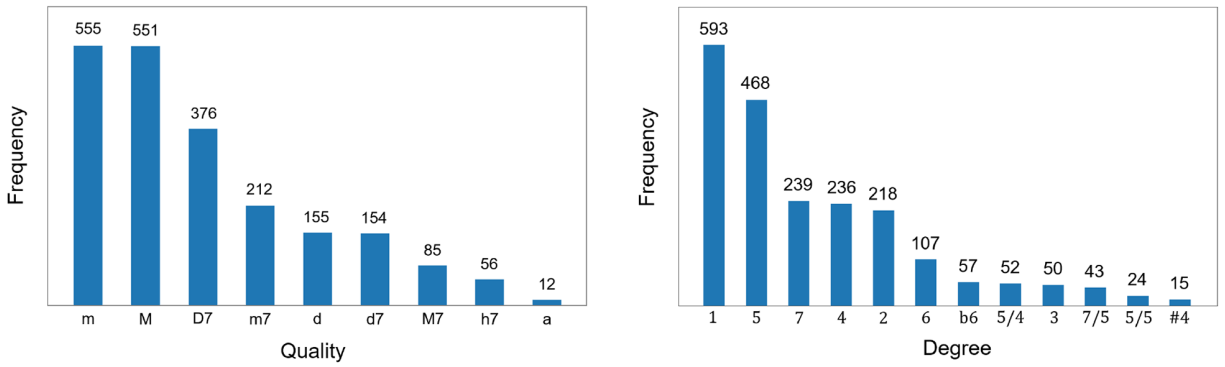
The musical pieces in the repertoire are represented as binary piano rolls with the time resolution of one 16th note, resulting in sequences of 88-dimensional feature vectors. A sliding window of length 128 (equal to 32 quarter notes) with a hop size of 16 is applied to the piano rolls to generate the instances for recognition. For the chord recognition task, we use the *maj-min* chord vocabulary (including 24 major and minor chords plus an additional 'others' class which is excluded from evaluation). The mapping of the chord qualities is shown in **Table 2**. We choose this vocabulary rather than other vocabularies of larger size for two reasons. First, both the BTC and the HT were evaluated using the same vocabulary. Second, there is still room for improvement in ACR even using this relatively small vocabulary. For the functional harmony recognition task, we decompose the RN labels into two parts, i.e., the key and the RN, whose vocabulary sizes are 42 and 5040 respectively, as delineated in **Table 3**.

3.3 Experimental Setting

We evaluate the BTC, the HT, and the HT* on the chord recognition and the functional harmony recognition tasks; and 4-fold cross validation is performed on each corpus. To create cross-validation sets, we naively assign a fold id to a piece according to the piece's id: fold_id = piece_id % 4. The training data are augmented via modulations (from 3 semitones down to 6 semitones up), leading to 10 times the original amount of data. As a result, the amounts of training and validation data are around (54320, 294)



(a) BPS-FH



(b) Bach Preludes

Figure 3: Statistics of the chord quality and degree annotations (some minor cases are omitted).

Table 3: Vocabularies of the functional harmony recognition task. The 21 tonics include {C,D,E,F,G,A,B} by {♮,♯,♭}; the 2 modes are {major, minor}; the 9 primary degrees are {1,2,3,4,5,6,7,b2,b7}; the 14 secondary degrees are {1,2,3,4,5,6,7,#1,#3,#4,b1,b3,b6,b7}; the 4 inversions are {root position,1st,2nd,3rd}.

Output	Component	Vocabulary Size
Key	21 tonics	42
	2 modes	
Roman Numeral	9 primary degrees	5040
	14 secondary degrees	
	10 qualities	
	4 inversions	

for the BPS-FH and (5380, 26) for the Bach Preludes. In addition, we train several variants of the BTC and the HT for the ablation study, and a CRNN for the comparison with the HT*:

- BTC-singleBi: the pair of uni-directional intra-MHAs is replaced with a single bi-directional intra-MHA.
- BTC-FC: the convolutional FFN is replaced by a fully-connected FFN.
- HT-noReg: the regionalization unit is removed.
- HT-noW: only the output of the final layer is used instead of the weighted sum of all the layers.
- CRNN: 10 one-dimensional convolution layers (convolving along the time dimension) with kernel size = 9 (equal to a window size of 2 quarter notes) plus 1 bi-directional Long Short-Term Memory (LSTM) layer.

The BTC-singleBi and the BTC-FC are employed to examine the effectiveness of the uni-directional intra-MHA pair and the convolutional FFN; the HT-noReg and the HT-noW are used to verify the contributions of the regionalization unit and the layer weights. As for the CRNN, we regard it as the benchmark for the HT* since they have a similar network topology. The architecture of the CRNN is modified from the network of Micchi et al. (2020), whose number of parameters is made comparable to the HT*. We set the number of repetitive layers to 2 for all models (other hyperparameters can be found in **Table 1**). During training, we set the dropout rate to 0.1; early stopping is applied once the model's performance stops improving on validation data for 10 consecutive epochs, and we report the best performance for evaluation.[2]

3.4 Evaluation Metrics

All models are evaluated in two aspects: 1) frame-wise chord recognition accuracy, 2) chord segmentation quality. The former shows the capability of a model to correctly predict chords at the frame level, while the latter assesses the predicted chord sequences from the perspective of chord segmentation. We utilize the directional Hamming distance (DHD) (Mauch and Dixon, 2010; Oudre et al., 2011) to evaluate the segmentation quality (SQ):

$$\begin{aligned} \mathrm{SQ} &= 1 - \max(\mathrm{DHD}(\mathbf{S},\hat{\mathbf{S}}),\mathrm{DHD}(\hat{\mathbf{S}},\mathbf{S})), \\ \mathrm{DHD}(\mathbf{S},\hat{\mathbf{S}}) &= \frac{\Sigma_{n=1}^{N}(|\mathbf{S}_n| - \max_{n'}|\mathbf{S}_n \cap \hat{\mathbf{S}}_{n'}|)}{\Sigma_{n=1}^{N}|\mathbf{S}_n|}, \end{aligned} \tag{5}$$

where $\mathbf{S}_n$ denotes the frames belonging to the nth segment of the annotated segmentation $\mathbf{S}$, and $\hat{\mathbf{S}}_{n'}$ denotes the frames belonging to the n'th segment of the predicted segmentation $\hat{\mathbf{S}}$. The SQ value reflects the similarity of two segmentations, ranging from 0 to 1. The higher the value, the better the segmentation quality. In particular, a value of 1 indicates that the two segmentations are exactly the same.

3.5 Results

3.5.1 Ablation Study and Comparison of BTC and HT

Table 4 shows the overall performance of each model. For the BTC-singleBi, using single bi-directional MHAs instead of uni-directional MHA pairs appears to lower the recognition accuracy when the amount of data increases (the case of the BPS-FH) and when the complexity of the task increases (the case of functional harmony). However, this may result from the fact that the number of parameters in the model is half of that in the BTC. For the BTC-FC, the substitution of the fully-connected FFN for the convolutional one is harmful to the performance in most of the cases, because a fully-connected network only computes the weighted sum of its inputs and does not take into account the temporally adjacent information which helps relate local features to higher-level semantics (Ren et al., 2019).

On the other hand, the HT surpasses the HT-noW in nearly all measures on the BPS-FH (while they are more or less comparable on the Bach Preludes), indicating that it might be beneficial to use information from all the layers of the network. Given that previous work has shown that different layers of a deep neural network tend to encode different types of information (Melamud et al., 2016; Belinkov et al., 2017), we believe that the employment of layer weights can increase the model's capability, especially when the encoder and the decoder of the HT are designated to different objectives (i.e., chord segmentation and chord recognition). Moreover, the HT outperforms the HT-noReg in four out of the five measures on the BPS-FH and in three measures on the Bach Preludes, validating the employment of the regionalization unit.

In comparison with the BTC, the HT appears to be more competent for it is more accurate in eight out of the ten measures (four from each corpus). In particular, the worst HT variant even outperforms all the BTC variants in terms of chord segmentation quality, showing that the concurrent estimation of harmonic changes benefits the outcome of chord segmentation.

Table 4: Evaluations with the BPS-FH dataset and the Bach Preludes. All the scores (in percentage) are averaged over 4 validation sets; the standard deviations of the scores are also provided.

BPS-FH					
Model	**Chord Symbol Recognition**		**Functional Harmony Recognition**		
	Accuracy	Segmentation	Key	Roman numeral	Segmentation
BTC	**82.46**$\pm_{1.55}$	**81.30**$\pm_{1.08}$	77.65$\pm_{1.83}$	**37.98**$\pm_{1.34}$	66.73$\pm_{4.05}$
BTC-singleBi	82.16$\pm_{1.66}$	80.78$\pm_{1.39}$	75.96$\pm_{0.79}$	35.77$\pm_{1.85}$	**68.83**$\pm_{1.69}$
BTC-FC	82.06$\pm_{1.83}$	81.24$\pm_{1.26}$	**78.40**$\pm_{2.10}$	37.60$\pm_{1.76}$	65.56$\pm_{3.86}$
HT	**83.19**$\pm_{1.65}$	**83.47**$\pm_{1.22}$	**77.94**$\pm_{2.24}$	**37.00**$\pm_{2.88}$	71.93$\pm_{2.72}$
HT-noW	83.06$\pm_{1.58}$	83.26$\pm_{0.71}$	77.13$\pm_{1.78}$	36.84$\pm_{2.39}$	**73.53**$\pm_{1.26}$
HT-noReg	83.19$\pm_{1.31}$	83.33$\pm_{1.26}$	76.70$\pm_{1.26}$	35.33$\pm_{1.79}$	70.51$\pm_{1.16}$
CRNN	79.79$\pm_{0.84}$	81.49$\pm_{1.91}$	75.56$\pm_{2.84}$	34.83$\pm_{1.38}$	67.75$\pm_{3.59}$
HT*	**83.98**$\pm_{1.08}$	**85.09**$\pm_{0.96}$	**79.07**$\pm_{2.70}$	**41.74**$\pm_{2.63}$	**75.50**$\pm_{1.72}$

Bach Preludes					
Model	**Chord Symbol Recognition**		**Functional Harmony Recognition**		
	Accuracy	Segmentation	Key	Roman numeral	Segmentation
BTC	74.12$\pm_{0.12}$	77.20$\pm_{3.64}$	**48.63**$\pm_{4.48}$	**25.25**$\pm_{1.76}$	**64.19**$\pm_{2.08}$
BTC-singleBi	**75.67**$\pm_{1.42}$	**78.85**$\pm_{4.81}$	46.24$\pm_{5.90}$	23.35$\pm_{1.99}$	60.40$\pm_{6.25}$
BTC-FC	75.53$\pm_{1.22}$	77.81$\pm_{4.51}$	46.05$\pm_{1.84}$	22.97$\pm_{2.19}$	57.24$\pm_{4.17}$
HT	**77.18**$\pm_{1.24}$	80.46$\pm_{3.36}$	**51.15**$\pm_{2.47}$	23.75$\pm_{2.20}$	66.82$\pm_{4.52}$
HT-noW	76.51$\pm_{1.45}$	**81.14**$\pm_{3.31}$	48.95$\pm_{2.88}$	**24.99**$\pm_{1.32}$	**67.61**$\pm_{4.75}$
HT-noReg	76.33$\pm_{1.23}$	80.76$\pm_{4.40}$	50.62$\pm_{3.93}$	23.82$\pm_{2.43}$	65.23$\pm_{4.80}$
CRNN	69.79$\pm_{1.15}$	79.47$\pm_{2.03}$	47.03$\pm_{6.59}$	18.53$\pm_{2.23}$	61.79$\pm_{1.83}$
HT*	**78.54**$\pm_{2.06}$	**83.86**$\pm_{2.24}$	**56.28**$\pm_{2.53}$	**25.95**$\pm_{1.67}$	**73.60**$\pm_{1.80}$

3.5.2 Improvement on the HT

It can be seen that the HT* substantially outperforms the CRNN benchmark in all cases. Although CRNN-based networks are known for their ability to jointly learn local features and model sequences, the employed CRNN fails to compete with the HT*. In comparison to the HT, the HT* obtains a consistent gain in overall performance, and the boost in segmentation quality is especially notable, validating the proposed improvement on learning the localized and position-related information. More strikingly, the HT* outperforms all the other models in comparison and sets new records for the current experiments. An examination into the multi-head attention mechanism reveals that the attention heads capture various concepts of musical harmony. As depicted in **Figure 4a**, two attention heads in the intra-MHA of the decoder display distinct attention patterns: one head (on the left hand side) appears to be aware of the chord regions, hence blocks of the attentive regions are formed along the diagonal of the attention map; the other head emphasizes more on the harmonic changes, resulting in several vertical lines traversing the attention map. We also observe different attention patterns in the inter-MHA of the decoder. As illustrated in **Figure 4b**, there are different block patterns signifying the harmonic structure of the input segment on the two attention maps. More specifically, the attention map on the left hand side reveals that the HT* is capable of recognizing the boundaries between the chords (as there are many darker lines around the positions where the chords change); while the attention map on the right hand side indicates that the tonic chord `(F:m)` and the dominant chord `(C:M)` put emphasis on different parts of the encoder sequence (see the red regions on the attention map).

To summarize, the performance of the Transformer-based model can be advanced by employing the intra-block intra-MHA and by enhancing the contextual information. Additionally, it is observed that the MHA units in the model can capture various harmonically meaningful characteristics of music, yielding prominent performances on both the chord recognition and the functional harmony recognition tasks. However, it should still be noted that these results are obtained by evaluating the model on only two piano solo corpora. Experiments using more diverse data are required for a more comprehensive assessment of the model.

4. Discussion and Future Work

The chord recognition accuracy is a simple and universal approach to evaluate ACR systems, but it is only part of the story. With formulating chord recognition as a classification problem, ACR systems using deep learning approaches are restricted to a *single* ground truth. In practice, however,

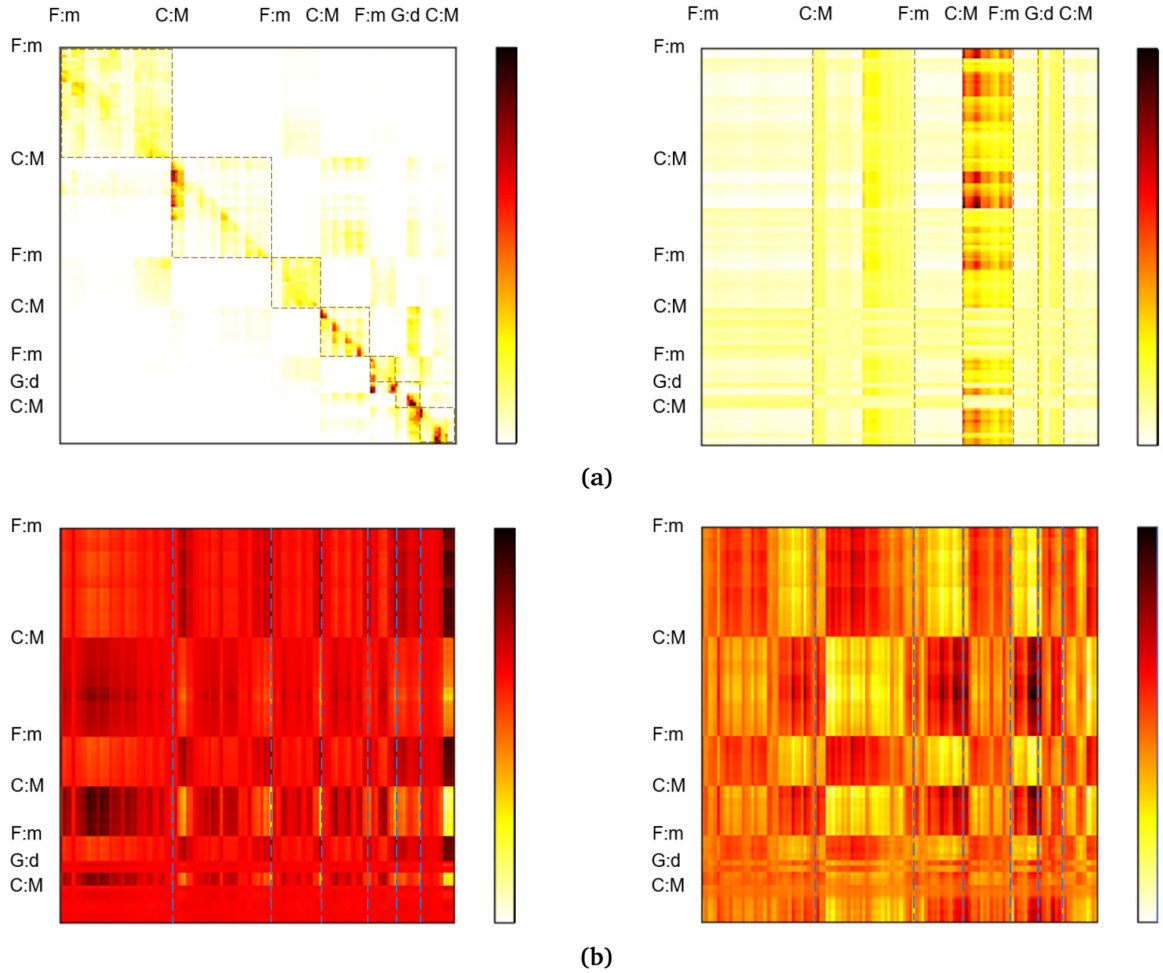


Figure 4: Examples of the attention maps in the MHA units; the color bars show the relative intensity of attention. The input segment is Beethoven's Piano Sonata No. 1, MM. 1–8. **(a)** Two attention heads in the intra-MHA of the decoder. The vertical and the horizontal axes respectively represent the queries and the keys, both of which indicate the same sequence to be recognized. **(b)** Two attention heads in the inter-MHA of the decoder. The vertical axis is the decoder sequence to be recognized (queries), and the horizontal axis is the encoder sequence (keys) for the chord change estimation (the positions where the chords change are indicated by vertical dashed lines).

several different annotations for the same harmonic entity are often equally viable due to the analytical essence of the recognition process, and the subjectivity in the ground truth will affect the evaluation of ACR systems (Ni et al., 2013). It is possible to take into account the relations between different annotations or predictions, and develop new loss functions for optimizing a deep learning model (Carsault et al., 2018).

Chord segmentation quality is another useful criterion for assessing frame-based ACR models, as the chord progressions obtained from these models should not be fragmented. Considering that chord boundary detection and chord recognition are intertwined problems, the ACR task may benefit from other advanced segmentation strategies, such as hierarchical representation (Chung et al., 2017), segmentation gates (Wang et al., 2018), and segment-directed attention (Hou et al., 2020). Instead of the frame-wise chord recognition, it is also worthwhile to explore methods for recognizing chords at a higher hierarchical level (Korzeniowski and Widmer, 2018).

As for functional harmony recognition, the experimental results suggest that there is still a long way to go to automatically produce RN analyses of sufficiently high quality. As functional harmony recognition relies heavily on the semantic content of music, it would be beneficial to leverage more high-level musical elements (e.g., metrical position and phrasing) for data representation. Moreover, designing reasonable output vocabularies is also important. Chen and Su (2019) predict the components of each RN label individually, while in the current experiments, they are combined into a single RN. The former reduces the output size of each component, but makes the components somewhat independent of each other; the latter alleviates the dependency issue but enlarges its output vocabulary size. Therefore, it is required to mediate between the two approaches. In addition, RN analysis is a fundamental tool for music theorists to uncover the tonal structure of music; hence human-in-the-loop approaches to functional harmony recognition are also valuable (Micchi et al., 2020).

Finally, currently available datasets for symbolic ACR are usually small and homogeneous. More work devoted to creation of symbolic datasets is thus welcomed. More diverse corpora would enable researchers to develop

standard benchmarks for a more comprehensive and systematic evaluation.

5. Conclusion

We systematically studied two Transformer-based ACR models in terms of model architecture, chord recognition accuracy, and chord segmentation quality. Furthermore, we also examined how the major components of the two models affect the overall performance. Based on one of the two models, we further improved the performance by leveraging the local context and the positional information of input music. In addition, experimental results showed that multi-head attention has the potential to capture various harmonically meaningful features in the scenario of ACR. We consider that attention-based models are promising for recognizing chords, not only because they have yielded fruitful results in various sequence modeling tasks, but also the attention mechanism is capable of accessing sequences in a relatively comprehensive way without being constrained by the receptive field of convolutions or by the sequential order of recurrences.

We have put emphasis on ACR in general and on the deep learning-based systems for symbolic music in particular, due to their potential to scale up the harmony-related annotation data. This enables us to investigate harmony from a macro perspective. With such a process, researchers will be able to provide a different insight into harmony than may be observed otherwise with a small number of musical instances. We hope that our research will draw more attention to symbolic ACR and encourage the MIR community to build datasets and develop techniques for symbolic music.

Notes

[1] The implementations of the BTC and the HT can be found at https://github.com/jayg996/BTC-ISMIR19 and at https://github.com/Tsung-Ping/Harmony-Transformer, respectively.

[2] The implementations of the models for evaluation can be found at https://github.com/Tsung-Ping/Harmony-Transformer-v2.

Acknowledgements

We thank the editor and the anonymous reviewers for their insightful comments.

Competing Interests

The authors have no competing interests to declare.

References

Ba, L. J., Kiros, R., and **Hinton, G. E.** (2016). Layer normalization. In *arXiv preprint arXiv: 1607.06450*.

Bahdanau, D., Cho, K., and **Bengio, Y.** (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and **Glass, J. R.** (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 861–872. DOI: https://doi.org/10.18653/v1/P17-1080

Boulanger-Lewandowski, N., Bengio, Y., and **Vincent, P.** (2013). Audio chord recognition with recurrent neural networks. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 335–340.

Carsault, T., Nika, J., and **Esling, P.** (2018). Using musical relationships between chord labels in automatic chord extraction tasks. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, (ISMIR)*, pages 18–25.

Chen, T. and **Su, L.** (2018). Functional harmony recognition of symbolic music data with multi-task recurrent neural networks. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 90–97.

Chen, T. and **Su, L.** (2019). Harmony Transformer: Incorporating chord segmentation into harmony recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 259–267.

Cho, T. and **Bello, J. P.** (2009). Real-time implementation of HMM-based chord estimation in music audio. In *Proceedings of the International Computer Music Conference (ICMC)*.

Chung, J., Ahn, S., and **Bengio, Y.** (2017). Hierarchical multiscale recurrent neural networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., and **Salakhutdinov, R.** (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 2978–2988. DOI: https://doi.org/10.18653/v1/P19-1285

de Haas, W. B., Magalhães, J. P., Veltkamp, R. C., and **Wiering, F.** (2011). HARMTRACE: Improving harmonic similarity estimation using functional harmony analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 67–72.

Degani, A., Dalai, M., Leonardi, R., and **Migliorati, P.** (2015). Harmonic change detection for musical chords segmentation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. DOI: https://doi.org/10.1109/ICME.2015.7177404

Degani, A., Dalai, M., Leonardi, R., and **Migliorati, P.** (2017). Audio chord estimation based on meter modeling and two-stage decoding. In *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 65–69. DOI: https://doi.org/10.1109/ISPA.2017.8073570

Deng, J. and **Kwok, Y.** (2016). A hybrid Gaussian-HMM-deep learning approach for automatic chord estimation with very large vocabulary. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 812–818.

Deng, J. and **Kwok, Y.** (2017). Large vocabulary automatic chord estimation with an even chance training scheme.

In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 531–536.

Devaney, J., Arthur, C., Condit-Schultz, N., and **Nisula, K.** (2015). Theme and variation encodings with Roman numerals (TAVERN): A new data set for symbolic music analysis. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 728–734.

Devlin, J., Chang, M., Lee, K., and **Toutanova, K.** (2019). 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 (NAACLHLT)*, pages 4171–4186.

Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., and **McAuley, J. J.** (2019). LakhNES: Improving multiinstrumental music generation with cross-domain pre-training. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 685–692.

Dong, H. and **Yang, Y.** (2018). Convolutional generative adversarial networks with binary neurons for polyphonic music generation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 190–196.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using Common Lisp Music. In *Proceedings of the International Computer Music Conference (ICMC)*.

Gotham, M. and **Ireland, M.** (2019). Taking form: A representation standard, conversion code, and example corpora for recording, visualizing, and studying analyses of musical form. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 693–699.

Gu, J., Bradbury, J., Xiong, C., Li, V. O. K., and **Socher, R.** (2018). Non-autoregressive neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Harte, C., Sandler, M., and **Gasser, M.** (2006). Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 21–26. DOI: https://doi.org/10.1145/1178723.1178727

He, K., Zhang, X., Ren, S., and **Sun, J.** (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and **Blunsom, P.** (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1693–1701.

Hori, T., Nakamura, K., and **Sagayama, S.** (2017). Music chord recognition from audio data using bidirectional encoder-decoder LSTMs. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1312–1315. DOI: https://doi.org/10.1109/APSIPA.2017.8282235

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., Laga, H., and **Bennamoun, M.** (2019). Bi-SAN-CAP: Bidirectional self-attention for image captioning. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. DOI: https://doi.org/10.1109/DICTA47822.2019.8946003

Hou, J., Guo, W., Song, Y., and **Dai, L.** (2020). Segment boundary detection directed attention for online end-to-end speech recognition. *EURASIP J. Audio, Speech and Music Processing*, 2020(1), 3. DOI: https://doi.org/10.1186/s13636-020-0170-z

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

Humphrey, E. J. and **Bello, J. P.** (2012). Rethinking automatic chord recognition with convolutional neural networks. In *Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA)*, pages 357–362. DOI: https://doi.org/10.1109/ICMLA.2012.220

Humphrey, E. J. and **Bello, J. P.** (2015). Four timely insights on automatic chord estimation. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 673–679.

Illescas, P. R., Rizo, D., and **Quereda, J. M. I.** (2007). Harmonic, melodic, and functional automatic analysis. In *Proceedings of the International Computer Music Conference (ICMC)*.

Jiang, J., Chen, K., Li, W., and **Xia, G.** (2019). Large-vocabulary chord transcription via chord structure decomposition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 644–651.

Korzeniowski, F. and **Widmer, G.** (2016). A fully convolutional deep auditory model for musical chord recognition. In *Proceedings of the 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. DOI: https://doi.org/10.1109/MLSP.2016.7738895

Korzeniowski, F. and **Widmer, G.** (2017). On the futility of learning complex frame-level language models for chord recognition. In *Proceedings of the AES International Conference on Semantic Audio*.

Korzeniowski, F. and **Widmer, G.** (2018). Improved chord recognition by combining duration and harmonic language models. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 10–17.

Lee, K. (2006). Automatic chord recognition from audio using enhanced pitch class profile. In *Proceedings of the International Computer Music Conference (ICMC)*.

Li, X. and **Wu, X.** (2015). Long Short-Term Memory based convolutional recurrent neural networks for large vocabulary speech recognition. In *Proceedings*

of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 3219–3223. DOI: https://doi.org/10.1109/ICASSP.2015.7178826

Lim, Y., Chan, C. S., and **Loo, F. Y.** (2020). Style-conditioned music generation. In *IEEE International Conference on Multimedia and Expo, ICME 2020, London, UK, July 6–10, 2020*, pages 1–6. DOI: https://doi.org/10.1109/ICME46284.2020.9102870

Masada, K. and **Bunescu, R. C.** (2017). Chord recognition in symbolic music using Semi-Markov Conditional Random Fields. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 272–278.

Masada, K. and **Bunescu, R. C.** (2019). Chord recognition in symbolic music: A segmental CRF model, segment-level features, and comparative evaluations on classical and popular music. *Trans. Int. Soc. Music. Inf. Retr.*, 2(1), 1–13. DOI: https://doi.org/10.5334/tismir.18

Mauch, M. and **Dixon, S.** (2010). Simultaneous estimation of chords and musical context from audio. *IEEE Trans. Audio, Speech & Language Processing (TASLP)*, 18(6), 1280–1289. DOI: https://doi.org/10.1109/TASL.2009.2032947

McFee, B. and **Bello, J. P.** (2017). Structured training for large-vocabulary chord recognition. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 188–194.

Melamud, O., Goldberger, J., and **Dagan, I.** (2016). Context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 51–61. DOI: https://doi.org/10.18653/v1/K16-1006

Micchi, G., Gotham, M., and **Giraud, M.** (2020). Not all roads lead to Rome: Pitch representation and model architecture for automatic harmonic analysis. *Trans. Int. Soc. Music. Inf. Retr.*, 3(1), 42–54. DOI: https://doi.org/10.5334/tismir.45

Miller, A. H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., and **Weston, J.** (2016). Key-value memory networks for directly reading documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1400– 409. DOI: https://doi.org/10.18653/v1/D16-1147

Neuwirth, M., Harasim, D., Moss, F. C., and **Rohrmeier, M.** (2018). The annotated Beethoven corpus (ABC): A dataset of harmonic analyses of all Beethoven string quartets. *Front. Digital Humanities*, 5. DOI: https://doi.org/10.3389/fdigh.2018.00016

Ni, Y., McVicar, M., Santos-Rodríguez, R., and **Bie, T. D.** (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE ACM Trans. Audio Speech Lang. Process.*, 21(12), 2607–2615. DOI: https://doi.org/10.1109/TASL.2013.2280218

Oudre, L., Févotte, C., and **Grenier, Y.** (2011). Probabilistic template-based chord recognition. *IEEE Trans. Audio, Speech & Language Processing (TASLP)*, 19(8), 2249–2259. DOI: https://doi.org/10.1109/TASL.2010.2098870

Parikh, A. P., Täckström, O., Das, D., and **Uszkoreit, J.** (2016). A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2249–2255. DOI: https://doi.org/10.18653/v1/D16-1244

Park, J., Choi, K., Jeon, S., Kim, D., and **Park, J.** (2019). A bi-directional Transformer for musical chord recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 620–627.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and **Tran, D.** (2018). Image Transformer. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4052–4061.

Passos, A. T., Sampaio, M., Kröger, P., and **de Cidra, G.** (2009). Functional harmonic analysis and computational musicology in Rameau. In *Proceedings of the 12th Brazilian Symposium on Computer Music (SBCM)*.

Pauwels, J., O'Hanlon, K., Gómez, E., and **Sandler, M. B.** (2019). 20 years of automatic chord recognition from audio. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 54–63.

Raphael, C. and **Stoddard, J.** (2004). Functional harmonic analysis using probabilistic models. *Computer Music Journal*, 28(3), 45–52. DOI: https://doi.org/10.1162/0148926041790676

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and **Liu, T.** (2019). Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3165–3174.

Rhodes, C., Lewis, D., and **Müllensiefen, D.** (2009). Bayesian model selection for harmonic labelling. In Klouche, T. and Noll, T., editors, *Mathematics and Computation in Music*, pages 107–116. Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-04579-0_11

Rocher, T., Robine, M., Hanna, P., and **Strandh, R.** (2009). Dynamic chord analysis for symbolic music. In *Proceedings of the 2009 International Computer Music Conference (ICMC)*.

Scholz, R. E. P. and **Ramalho, G. L.** (2008). COCHONUT: recognizing complex chords from MIDI guitar sequences. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 27–32.

Shaw, P., Uszkoreit, J., and **Vaswani, A.** (2018). Selfattention with relative position representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT)*, pages 464–468. DOI: https://doi.org/10.18653/v1/N18-2074

Sheh, A. and **Ellis, D. P. W.** (2003). Chord segmentation and recognition using EM-trained Hidden Markov

Models. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*.

Shen, T., Zhou, T., Long, G., Jiang, J., and **Zhang, C.** (2018). Bi-directional block self-attention for fast and memory-efficient sequence modeling. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Stark, A. M. and **Plumbley, M. D.** (2009). Real-time chord recognition for live performance. In *Proceedings of the International Computer Music Conference (ICMC)*.

Tsui, V. and **MacLean, W. J.** (2002). Harmonic analysis using neural networks. In *Proceedings of the International Computer Music Conference (ICMC)*.

Tymoczko, D., Gotham, M., Cuthbert, M. S., and **Ariza, C.** (2019). The Romantext format: A flexible and standard method for representing Roman numerial analyses. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 123–129.

Ueda, Y., Uchiyama, Y., Nishimoto, T., Ono, N., and **Sagayama, S.** (2010). HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5518–5521. DOI: https://doi.org/10.1109/ICASSP.2010.5495218

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and **Polosukhin, I.** (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Wang, Y., Lee, H., and **Lee, L.** (2018). Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6269–6273. DOI: https://doi.org/10.1109/ICASSP.2018.8462002

Yang, M., Su, L., and **Yang, Y.** (2016). Highlighting root notes in chord recognition using cepstral features and multi-task learning. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–8. DOI: https://doi.org/10.1109/APSIPA.2016.7820865

Yoshioka, T., Kitahara, T., Komatani, K., Ogata, T., and **Okuno, H. G.** (2004). Automatic chord transcription with concurrent recognition of chord symbols and boundaries. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*.

Zenz, V. and **Rauber, A.** (2007). Automatic chord detection incorporating beat and key detection. In *Proceedings of the IEEE International Conference on Signal Processing and Communications (ICSPC)*, pages 1175–1178. DOI: https://doi.org/10.1109/ICSPC.2007.4728534

Zhou, X. and **Lerch, A.** (2015). Chord detection using deep learning. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 52–58.

How to cite this article: Chen, T.-P., and Su, L. (2021). Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 1–13. DOI: https://doi.org/10.5334/tismir.65

Submitted: 10 May 2020 **Accepted:** 07 January 2021 **Published:** 24 February 2021