

Computational Music: Analysis of Music Forms

Jing Zhao[1], KokSheik Wong[1(✉)], Vishnu Monn Baskaran[1], Kiki Adhinugraha[3], and David Taniar[2]

[1] Monash University Malaysia, Bandar Sunway, SL 47500, Malaysia
{jing.zhao,wong.koksheik,vishnu.monn}@monash.edu
[2] Monash University, Clayton, VIC 3800, Australia
david.taniar@monash.edu
[3] La Trobe University, Bundoora, VIC 3086, Australia
k.adhinugraha@latrobe.edu.au

Abstract. With the development of computational science, many fields, including computational linguistics (sequence processing) and computational vision (image processing), have enabled various applications and automation with satisfactory results. However, the development of Computational Music Analysis (CMA) is still in its infancy. The main factor hindering the development of CMA is the complex form found in music pieces, which can be studied and analyzed in many different ways. Considering the advantages of Deep Learning (DL), this paper envisions a methodology for using DL to promote the development of Music Form Analysis (MFA). First, we review some common music forms and emphasize the significance and complexity of music forms. Next, we overview the CMA in two different processing ways, i.e., sequence-based processing and image-based processing. We then revisit the aims of CMA and propose the analysis principles that need to be satisfied for achieving the new aims during music analysis, including MFA. Subsequently, we use the fugue form as an example to verify the feasibility and potential of our envisioned methodology. The results validate the potential of using DL to obtain better MFA results. Finally, the problems and challenges of applying DL in MFA are identified and concluded into two categories, namely, the music and the non-music category, for future studies.

Keywords: Music form · Computational music analysis · Deep learning · Computational science

1 Introduction

Music is an ancient yet common information carrier used from generation to generation. As the soul and essence of music, *music form* (MF) organizes individual musical notes together to form layered and expressive sound. Hence, many studies, applications, and analyses of music are inseparable from MF. These include music education, musicology, and music appreciation, which uses MF directly. In

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13956, pp. 366–384, 2023.
https://doi.org/10.1007/978-3-031-36805-9_25

addition, other applications emerge from utilizing the derived information from MF, such as music composition [20,68], music emotion classification [51], genre classification [17], thumbnail generation [46], making chord and downbeat estimation [27], music recommendation [58], to name a few. Furthermore, music has also been used for health and information security purposes in recent years. For example, music has been adopted in psychology [37], Covid-19 treatment [39], heart disease treatment [16], neurology [25,60], and cryptography [38,43]. Therefore, MF has become the development cornerstone that provides a basis and important music characteristics for in-depth research in many fields.

Popular music form usually lacks richness with just a few variations (e.g., key, rhythm, chord) [70]. In contrast, classical music is of complex forms, and it is informative with deep content. In fact, the notion of MF is more applicable to the case of complex classical music [18], which has different forms such as *theme and variation*, *fugue*, *sonata*, etc. Interestingly, research also shows that classical music is closely related to human physiological responses, such as changes in mood, heartbeat, etc. [70]. Therefore, regardless of research significance or traditional MF definition in musicology, classical music form is worthy of attention.

While some domains have advanced significantly using deep learning models (e.g., classification for MS-COCO @ 65.4% of mean average precision [65] using CNN model, and IMDB Review data set @ 96.21% of accuracy [55] using XLNet model), we envision that music form analysis (MFA), which is an important sub-field in computational music analysis (CMA), can also benefit from the data and computationally driven approaches (e.g., deep learning models). Therefore, being the first article that envisions the analysis of the complex and informative classical music form based on the development of the data-driven approach, we set to explore a few interesting questions, including:

Q1: What is music form, and why is it complex yet important to achieve advanced results?
Q2: What is CMA?
Q3: Why is the development of CMA slow, and how can the CMA develop better?
Q4: What is a data-driven approach, and how can it work for MFA?

The rest of the article is structured as follows. Section 2 addresses the first question (Q1). Specifically, we overview the classical music forms and emphasize the significance of MF in music and other fields as well as the complexity of MF. Such discussions are conducted from two processed ways, namely, sequence-based processing and image-based processing. In Sect. 3, we review and clarify the eventual aims of the whole CMA field. We also stipulate the principle to which the results from the sub-field of CMA (e.g., MFA) should conform for a deeper study. Such discussion address question Q2 and Q3. Since many computational science studies have benefited from the development of data-driven approach, we introduce a data-driven approach for MFA in Sect. 4.1 and Sect. 4.2, which addresses question Q4. To demonstrate the feasibility of applying the DL methodology to MFA, we conduct an initial experiment for fugue form analysis and present the results of the case study in Sect. 4.3. Finally, Sect. 5 identifies the challenges of MFA and concludes this paper.

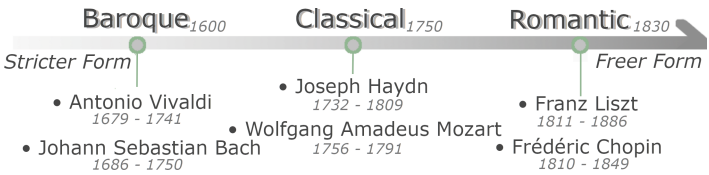


Fig. 1. Three eras of classical music and some representative composers.

2 Music Form

In this section, we first overview some common music forms in classical music and its three eras. Subsequently, we systematically highlight the significance of *music form* in music and other domains and analyze its complexity.

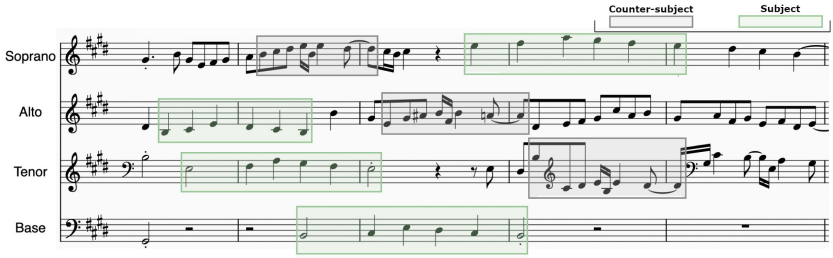
Typically, the term *music form* (MF) refers to the classical music form rather than the popular music form [18]. The first reason is that the form of popular music is relatively simple, with merely a few variations (e.g., key, rhythm, chord) [70]. Secondly, the components of different pop music are similar, including verse, chorus, bridge, etc. Under this situation, although the simple form satisfies the aesthetics of the general public, the music it presents lacks richness and depth. Therefore, complex classical music is often considered in the form analysis under traditional musicology. Furthermore, interestingly, research shows that classical music, which contains deep content while being expressive, is closely related to human physiological responses, such as changes in mood, heartbeat, etc. [70]. Therefore, regardless of research significance or the traditional definition of MF in musicology, classical music form is worthy of attention.

Classical music can typically be divided into three eras, namely, *Baroque*, *Classical*, *Romantic*, as shown in Fig. 1. Furthermore, the music in the *Baroque* era has a relatively strict form. As time passes, the meaning of the music form is weakened to a certain extent, noticeably in the *Romantic* era [66]. In other words, the composition of music becomes relatively more free. The common classical music forms include *fugue*, *theme and variation*, *sonata*, to name a few. We will discuss these music forms based on their most prevalent era.

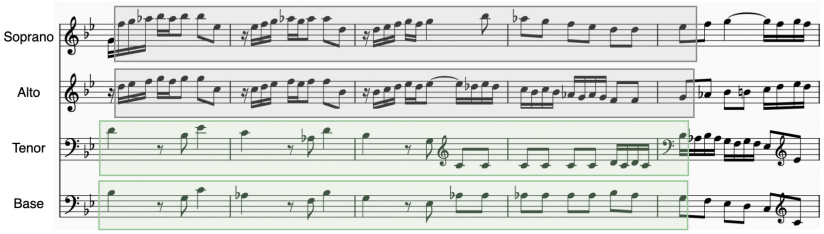
2.1 Fugue

The *fugue* form prevailed in the Baroque period (the 1600 s)). It is polyphonic with a strictly ordered counterpoint form [59,61]. In other words, the complexity and richness of this music are attributed to stacking voices. For most cases, a *fugue* contains four voices, namely, *Soprano*, *Alto*, *Tenor*, and *Bass* [19], as shown in Fig. 2. In addition, Fig. 2 also shows the annotation of the *Subject* (S) and *Counter-Subject* (CS), which are the main components in the *fugue* form. Generally speaking, S is the main melody, and all other melodies are composed around S.

Technically, S and CS have different characteristics from the perspective of music composition. First, S will start at the beginning of the *fugue* music



(a) A segment of J.S. Bach's BWV878



(b) A segment of J.S. Bach's BWV885

Fig. 2. Two segments showing the *fugue* form with four voices.

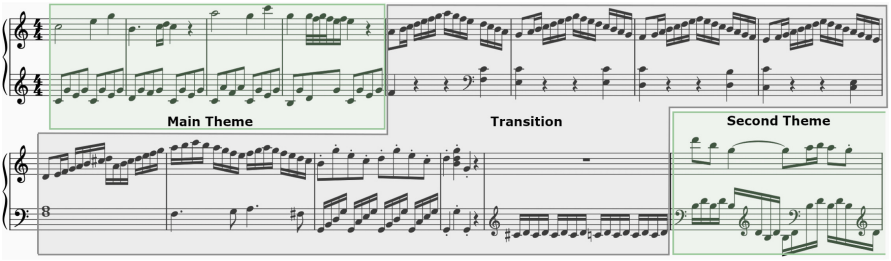


Fig. 3. An example of the *sonata* form's *exposition* part (Mozart K545).

independently (i.e., in a single voice). In addition, different S can appear in different voices and overlap in the temporal dimension, such as the two examples shown in Fig. 2. Such overlapping feature is the common composition technique in the *fugue* form, viz., *stretto* [59]. This technique is most used at the end of the music to add a sense of compactness to the music [59]. For CS, it is often accompanied by S, and sometimes, CS will continue following S (see Fig. 2) [41, 56]. Additionally, in the whole *fugue* form, the S and CS are repeated melody segments with artful variations. The common variations are tonal, duration, pitch interval, note augmentation or diminution [26,31,53].

2.2 Sonata

Sonata form becomes noticeable since the Classical era [4,57]. Such form is large-scale with three main parts at the *structure level*, namely, *exposition*, *develop-*

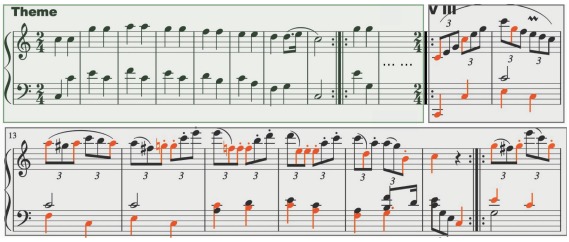


Fig. 4. The segments of a *theme* melody and the third *variation* (VIII) melody in Mozart's KV265 music (*theme and variation* form). The notes in the VIII that are similar to the *theme* melody are highlighted in orange. (Color figure online)

ment, and *recapitulation* [2,7]. These three parts shape the whole piece of music. Furthermore, sometimes, the *sub-structure level* of the sonata's *exposition* and *recapitulation* parts can still be divided further. Figure 3 shows an example of the *sonata* form with two themes in the *exposition*. By adding a new theme or key transition, the *sonata* form can express more conflict and tension in emotion [48].

2.3 Theme and Variation Music

The *theme and variation* (TV) form is another important music form in the classical era [52]. At the *structure level* (i.e., highest level), the TV form consists of a *theme* segment (T) at the beginning of the music and a series of *variation* segments (V). Different V segments innovate the T segment from different aspects [48]. In other words, the V segments will be composed based on T with some changes, such as notes number change, key variation, etc., so that the V segments are similar to T to a certain extent. An example of the T and a V segment is shown in Fig. 4. Here, the notes in the third V segment (labeled as VIII) that are similar to the T melody are highlighted. Furthermore, in the segment labeled VIII, the main change in the treble part ($\mathbf{\mathcal{G}}$) is notes diminution and adding more transition notes, while the main change in the bass part ($\mathbf{\mathcal{F}}$) is a variation of octave number with a similar contour. However, in the TV form, the keys of most V segments are the same as the key of the corresponding T segment. These are illustrated in Fig. 9. The same argument applies to the number of bars in the segments. Hence, the TV form is more concise in comparison to the *sonata* form [13].

2.4 Other Forms

In addition to the above three common yet complex music forms, there are other music forms. For example, the *binary* form and the *ternary* form include two and three sections/themes, respectively [23]. Among these sections/themes, they may involve key transitions (e.g., a major key transition to the dominant key) [23]. The typical example of the *binary* form is J.S. Bach BWV996, which contains

only two sections. Interestingly, after expanding these basic forms, such as adding more repetitions of the sections/themes, the *Rondo* form is constructed. Alternatively, when adding more repetitions and variations, the *theme and variation* form is constructed. Hence, the *binary* and *ternary* forms are arguably the simplest and most basic forms.

2.5 Significance and Complexity

After understanding some common music forms, we try answering the question: *"What is music form, and why is it complex yet important to achieve advanced results?"*. Specifically, the main music elements are notes [32], which can exist objectively and independently without any form. After structuring notes into a series of segments, a layered and expressive music piece is formed. Hence, *music form* (MF) is the soul of music, and music cannot exist without a form. Due to the importance and indispensability of MF in music, the study of MF benefits many research fields, including music education, musicology, and music appreciation. Some derivative fields rely on MF information as well, and its roles are broadly summarized below:

1. Providing new research methodology (e.g., different rules for notes combination in different music forms) for the field of arrhythmia research [16] and cryptography [38], etc.
2. Adding emotional value to music [51], and in a way participating in the treatment of Covid-19 disease [39].
3. Serving as the basic features for genre classification tasks [17], music attribute calculation [27], music recommendation algorithm [58], structured music composition [20,68] to low the threshold of composition for non-professionals [34], neurological treatment (movement disorders [25], aphasia [60], and psychiatric [11]).
4. Serving as a key feature of music visualization [46], which is further enhanced for music education and appreciation [6,12,22], music summarization and retrieval (e.g., cover song retrieval [21], music thumbnail [46]), music cryptography [43], and other fields.
5. Contributing to treating personality disorder patients as a social bond [37].

Although MF enables many applications, it is difficult to analyze because of its complexity, thus inhibiting it from unleashing its full potential. Specifically, MF is not a simple mechanical combination of segments. It manages the melody from two music perspectives of structure (i.e., the rhythm and pitch structures) [32]. The rhythm and pitch are responsible, respectively, for the rhythmic and pitch relationship (i.e., tonality and harmony) among notes. Although these two structures are relatively independent and can be discussed separately [32], without any of such two structures, the music becomes monotonous and poor in emotions and content, thus losing the meaning of MFA. Hence, the informative classical music form always involves these two perspectives of structure.

The other factor contributing to the complexity of MF is the rich and robust relationships of music. The relationships within the music form can be roughly

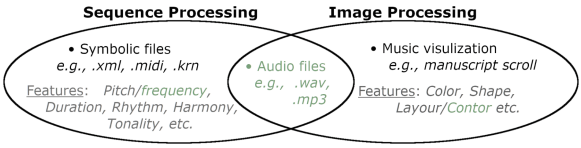


Fig. 5. Two common music processing ways and available features.

divided into two categories, internal relations (i.e., relationships of notes within a segment) and external relations (i.e., relationships among segments). Such rich relationships are the premise of music being expressive and capable of carrying in-depth content, but unfortunately, the rich relationships also complicate the MFA task.

3 Computational Music Analysis

This section reviews computational music analysis (CMA), contrasting it with computational vision and linguistics. We clarify the new aims of CMA and promote its development, focusing on two principles obeyed by the CMA results.

As a product of human beings, music has established a close connection with people from multiple dimensions (e.g., business, culture, and life) in multiple ways (audio, image, and symbolic files) [10,50]. In the early years, music analysis was performed manually by humans. Thanks to the recent advancement in computational power and machine learning capabilities, CMA has received some attention and emerges as a rising field. Typically, CMA can be considered a sequence processing task (e,g., symbolic files), or an image processing (e.g., manuscript scroll) task, as shown in Fig. 5. Unlike the symbolic file and manuscript scroll, which fall clearly in the respective categories, music presented in the audio form is more interesting. Specifically, when managing the music audio as a sequential signal, CMA can be treated as computational linguistics (sequence processing). On the other hand, when visualizing the music audio as a spectrogram, it can be treated as computational vision (image processing).

However, although CMA may adopt the existing information processing methodologies designed for other computational sciences, such as sequence-based computational linguistics [42] and image-based computational vision [44,62], CMA is still an emerging interdisciplinary field because of the uniqueness of music.

Like text for natural languages, music has a context with a temporal structure. However, unlike text for natural languages that assume general features in language sequences (e.g., collocations, logical vocabulary, negative words), there are no general features in the sequence of music. On the other hand, from the perspective of images, the basic elements are color and shape. Such basic elements can be used to distinguish content in different images (e.g., recognizing document layout [63,71] via the basic spatial elements). Although a music manuscript scroll also contains black-and-white colors and note shapes, the melodic content of the

notation is recorded by the position of the note (relative to the 5-lines) and the shape of the notes together, but not the colors and shapes themselves. Hence, despite the similarities, it is not straightforward to directly adopt methodologies designed for other matured fields to perform CMA due to the aforementioned intrinsic differences.

3.1 Aims and Principles

It is generally perceived that CMA mainly aims to produce interesting music results or to assist human analysts [3], hence CMA's results should conform to human cognition of music. Based on such perception and aim, the CMA results pay less attention to whether they can be reused in other fields, such as MIR, which seriously hinders the development of CMA and reduces the value of CMA results. Hence, with technological improvement and the rising trend of interdisciplinary research, it is necessary and timely to re-visit the perception and aim of CMA [45].

Referring to other matured fields in computational science and putting aside technological limitations and complexity, we conjecture that CMA can have a higher purpose, namely, understanding music more intelligently and professionally and even replacing human analysis in the future.

However, to achieve these purposes, the current principle adopted (viz., whether the results meet the human cognition of music) to obtain CMA results is insufficient. Based on our analysis, the CMA's results should satisfy two principles: 1) conforming to human cognition of music, and 2) being reusable in other research investigations.

Firstly, conforming to the human cognition of music can ensure that the CMA's results are trustworthy and authoritative from the music perspective. However, it is impossible to conform absolutely to the cognition principle, as argued by Marsden et al. [45]. Hence, for different studies, the researcher needs to decide on the level to which this principle is conformed.

Secondly, the principle stipulating that computers can understand and process CMA results ensures the usability of the results. This leads to two issues; one is the form of the result, and the other is the objectivity of the result. For the former issue, the form of the result needs to be catered to the computer instead of only conforming to the form of human vision or hearing. For the latter issue, although absolute neutrality or objectivity is also unrealistic [3] and CMA usually relies on some preliminary human knowledge [45], the result of CMA cannot depend excessively on prior knowledge of human cognition of music. Results involving too much subjective knowledge will include prejudice, thus affecting the subsequent application and reducing the value and authority of the CMA results. When the two aforementioned principles are satisfied simultaneously, CMA can provide more authoritative and reusable results.

In fact, no clear indicators can measure whether the results of CMA are in line with human cognition of music or whether the results of CMA are objective enough [45]. However, it is necessary to follow the guidance of the above

principles for highlighting the value of CMA (including MFA) for further study in many fields.

4 Envisioned Methodology

Based on the above analysis, the MFA methods have two processing ways (i.e., sequence and image). So far, for image processing (e.g., the visualization of audio files), some methods analyze the structure by manually labelling [15,49], searching for mechanically repeated melody segments [14,67], or detecting short segments and merging them to meet the length of the music form level [9,35]. Even when using deep learning (DL) techniques to detect the components of MF, such as Deep Belief Network [8] or Convolutional Neural Network [30], the results of music segmentation are subpar (i.e., 40%-50% of accuracy).

Relatively speaking, sequence-based methods have received less attention than image-based methods. Nonetheless, sequence-based methods have focused on the analysis of complex classical music forms, such as *fugue* form [28,29], *sonata* form [2,7]. Their main research idea is to find repetitive segments through a dynamic programming algorithm. Then, they use the Viterbi algorithm and Hidden Markov Model (HMM) to calculate the structure of the MF (i.e., the label of detected segments). Indeed, these methods contribute more in the MFA fields, but they have lower robustness and more limitations. For example, the *fugue* form analysis methods proposed by Giraud et al. [28,29] require three assumptions, including: (as1) the length of S, and (as2) the position of CS, and (as3) using the voice-separated symbolic files to simplify the parallel issues (e.g., multiple voices) in the *fugue*.

In short, there is still much room for improvement in both sequence-based and image-based MFA methods to serve the newly identified purpose. Therefore, considering that DL technology has been widely used in recent years and has achieved commanding results in many fields, we explore whether DL can break through the status quo of MFA for complex music forms.

4.1 Feature Engineering

In the DL era, common feature engineering includes handcraft and non-handcraft techniques. As the name implies, handcraft feature engineering artificially selects music features and embeds them in neural network models. In contrast, non-handcrafted feature engineering automatically extracts music features directly by using some machine-learning techniques. However, although the non-handcrafted feature engineering techniques are automatic with fewer answer leak risks, their performance is not always better than that of handcrafted feature engineering techniques [54]. Therefore, handcraft feature selection is still common in the field of music analysis. The common features used in handcraft feature engineering for CMA are shown in Fig. 5.

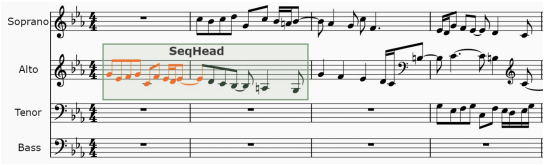


Fig. 6. An example showing the data description with *SeqHead* proposed for the MTBert model. The notes that are highlighted are the ground truth *subject* melody.

4.2 Deep Learning

The neural network model of deep learning simulates the biological nervous system in the human brain [1]. A large number of units in neural networks and brain neurons have similar functions. They can amplify some basic features to multiple layers, thus capturing more details [40]. In recent years, the neural network has already become a powerful computing approach thanks to the availability of data in large volumes. This ability enables DL to obtain state-of-art results (e.g., MF-COCO @ 65.4% of mean average precision [65] using CNN model, and IMDB Review data set @ 96.21% of accuracy [55] using XLNet model). The popular deep learning models include RNN, BERT, and CNN, to name a few. We will introduce them from the perspective of sequence processing (RNN, BERT) and image processing (CNN).

Specifically, for sequence processing, compared to the RNN-based model that learns sequentially and is unable to handle long-term relationships between contexts [5,33], BERT is better at processing information in the temporal dimension with rich long-term relationships. The advantage of the transformer-based model (e.g., BERT model [24]) is attributed to the position encoding and the self-attention mechanism. Such two mechanisms enable the model to permute the position of the input sequences to learn the relationship among contexts. In addition, they can better deal with contextual information with long-term relationships [69]. In short, a large-scale transformer-based model has a high capability to learn context and long-term relationships and has stronger interdisciplinary adaptability [36].

On the other hand, the CNN model is typically utilized to learn from images, which contain spatial information. This is due to the fixed-size kernel of CNN, which performs convolution operations on multiple layers. In other words, for an image, its local feature is meaningful, and the relationship among long-distanced elements is not strong. Therefore, the CNN model can capture many local features [5]. However, the local features of the musical image (e.g., manuscript scroll) may be meaningless or unimportant. For example, there may be a large amount of transitional melody in two *subjects* in a *fugue* form. In this case, the limited kernel size will lose or ignore some important information in the long-term relationship. Therefore, CNN is better at capturing short-term context relationships [64], but is not so attractive for long-term context relationships.

Inspired by the BERT model's ability to handle contextual information with long-term relationships, we conduct preliminary experiments using a proposed BERT-based model to explore the feasibility of applying DL for the purposes of MFA.

4.3 Case Study: Fugue Form Analysis

In this case study, we try weakening the assumptions (i.e., as1, as2 and as3 in Sect. 4) as a means to fulfill the second principle (viz., regarding reusability in Sect. 3.1). Note that, in the same MF, there is no clear relationship and general features between different pieces of music. To enhance the connection among the different *fugue* music files, which in turn, helps the DL model to learn, we propose a general feature to describe each piece of music, namely *SeqHead*. As shown in Fig. 6, *SeqHead* contains the first S occurrence of a *fugue* that starts independently. The length of *SeqHead* refers to Giraud et al. [28]'s statistical results (d_I), that is, $L_m + max(d_I), d_I \in [-8, 6]$, where L_m is the length of the first voice before the second voice, and d_I is the difference between L_m and the correct length of S in J.S. Bach's Well-Tempered Clavier (WTC) book I data set with 24 symbolic voice-separated fugue files [28]. After extracting *SeqHead* for each music piece, the music will be separated into segments with a length of 64 notes and the *SeqHead* of each music piece is concatenated to every segment of the same music.

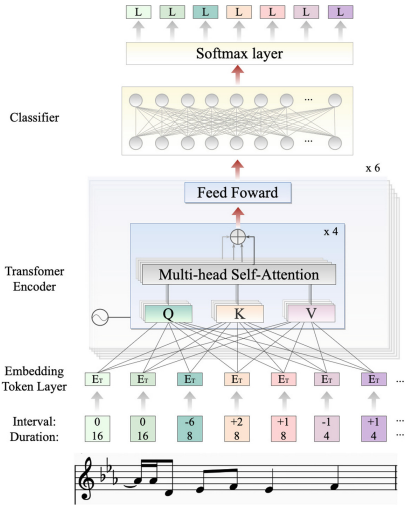


Fig. 7. The overall architecture of the MTBert model.

Next, we design a BERT-based model called *Music Tiny BERT model (MTBert)* for *fugue* form analysis, viz., extracting the S and CS in a given *fugue* form. An overview of the architecture of the proposed MTBert is shown in Fig. 7. Meanwhile, based on our designed data description method, we use

LSTM as the baseline model for comparison. Both models are trained and validated using WTC book I data set.

Table 1 shows the results of the note-level binary classification (i.e., S/non-S and CS/non-CS) obtained for BWV869 in the WTC data set using LSTM and MTBert. In the case of detecting S, the performance of LSTM is 73.3% for F1 score with high recall (91.7%). In the case of detecting CS, despite attaining 100% recall, LSTM predicts many False Positive errors causing extremely low precision. Hence, LSTM is non-reliable and infeasible to detect CS. On the other hand, using the same setup (*SeqHead*) and assumption (voice-separated input files) with as LSTM, MTBert can detect S with high precision (97.0%) and recall (90.6%). Furthermore, without making any assumption on CS, MTBert can detect more than 50% of the notes in CS with more balance recall and precision than LSTM. Note that such performances are better than those reported by Giraud et al. [28] who focus more on occurrence detection[1] based on the assumptions of as1 for S, as2 for CS, and as3 for the input files. However, under enough assumption, Giraud et al.'s [28] method cannot detect any occurrences of CS in BWV869. Although Giraud et al. [28] can detect all occurrences of S, there are still many errors in terms of length.

Table 1. The evaluation of MTBert and the baseline model of BWV869

	S			**CS**		
	Recall	Precision	F1 score	Recall	Precision	F1 score
LSTM	91.7%	61.1%	73.3%	100.0%	16.3%	28.0%
MTBert	97.0%	90.6%	93.7%	56.4%	74.8%	64.3%

Since the current techniques for computational *fugue* form are very limited, it is impossible to achieve significant improvement overnight. However, our method could at least demonstrate that it is possible to improve the performance using DL method under fewer assumptions, in comparison to the results obtained by non-DL methods. Furthermore, unlike the detection of occurrence allowing errors of length, MTBert's results are clear with zero-tolerance in length so that the results can be reused directly for other tasks. Hence, such outcomes follow the principle (i.e., the second principle involving objectivity and reusability in Sect. 3.1) of CMA analysis. Overall, the results of our case study, which is the first of its kind, can verify the feasibility of the application of DL on MFA and the potential of reaching the new purpose of MFA using DL.

[1] The mismatch in lengths for S and CS would not affect the results on the detected occurrences.

5 Problems and Challenges

Despite the promising outcome report above, applying DL on MFA still faces some challenges. We divide the challenges into two broad categories, namely, the music category and the non-music category. Here, the music category refers to the challenge caused by the complexity of the MF itself and the features of processing different music carriers (viz., sequence-based and image-based processing). On the other hand, the non-music category refers to the challenge of using DL on the complex MFA task.

5.1 Music Category

Under the music category, there are four subchallenges. The first subchallenge is the nonexistence of a general characteristic in different MFs. In other words, for the features shown in Fig. 5, they are utilized only to record the music notes instead of any other information regarding the MF. For example, no melody contour (pitch interval combination), fixed pitch/rhythm/harmony combination, color, or shape can represent or reflect the characteristics of a certain MF. This subchallenge is the reason why some researchers believe that there are no general features for creating a general method that can accommodate all music forms [47,67].

The second subchallenge is that some variation features of the MF's components (i.e., rhythm and pitch) cannot be computed directly. For example, in the *fugue* form shown in Fig. 2, the S or CS repeats with some variations, such as notes diminution (rhythm perspective), key variation (pitch perspective), etc.

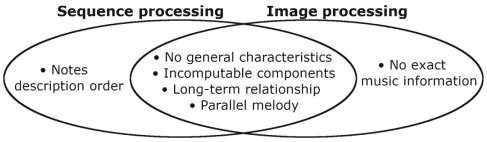


Fig. 8. Some challenges in the music category.

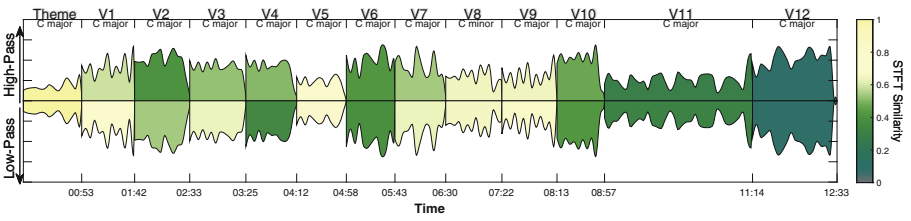


Fig. 9. Visualization of a *theme and variation* form (Mozart's KV265). V refers to the *variation* and the color shows the level of similarity between every V segments and the *theme* from the high-pass part and the low-pass part. These two parts can be understood simply as the treble part and the bass part.

Therefore, these changes take place in different music perspectives, and they cannot be quantified directly from the available features shown in Fig. 5.

The third subchallenge is that the internal relationships of music (i.e., the connection of different segments) are strong and long-term. A typical example is that every *variation* segment in *theme and variation* form is based on the *theme* segment, so there are strong relationships among the *variation* and *theme* segments. Figure 9 is a visualization of a *theme and variation* form in Mozart KV265. In this figure, the time distance between the last *variation* segment and the *theme* segment is around 11 min. This duration is too long for both the sequence processing and image processing approaches.

The last subchallenge is the parallel melody in music. A typical example is shown in the segments of the *fugue* in Fig. 2. As highlighted in green boxes, the components of this form (i.e., S and CS) can appear simultaneously. Hence, the boundary of music form components will overlap. Similarly, the *chords* in music with multiple notes at an instant also involve the problem of melody parallelism.

In addition to the four subchallenges mentioned above, analyzing MF using sequence or image-based methods also has a challenge each. For the sequence processing method, the order of the notes will be sensitive. Hence, the order to process the inseparable notes in parallel melody needs to be addressed. For the image processing method, even if one visualizes the image from the audio files or the manuscript scroll score, no exact music notes information can be used. Hence, the results of MFA are prone to bias.

5.2 Non-music Category

Under the non-music category, the vague definition of MF is an obstacle for training DL models. In other words, there is no clear definition of the corresponding segments in the higher/lower-levels within a piece of music. Hence, different definitions of the boundaries lead to different training results. Normally, when the segments in the music cannot be aggregated further, it is considered to be the highest-level music form, and comparatively, the highest level of form is less controversial.

Furthermore, due to the complexity of the music form, an approach needs to be designed for every specific music form [47,67]. This leads to a situation where there are many datasets of classical music, but only a limited number of them could be used for MFA. In other words, the usable dataset for every specific music form will be limited. Moreover, using the DL technique on a small-scale dataset is prone to suffer from the overfitting problem, i.e., learning features with low generalizability.

In short, the MFA task becomes very challenging due to the complexity of the music form. Although DL has achieved good results in various fields, the application in MFA tasks still faces some fundamental challenges, which need to be addressed.

6 Conclusion and Outlook

Music Form Analysis (MFA) is one of the most important sub-fields in Computational Music Analysis (CMA). Two typical ways of MFA are sequenced-based methods and image-based methods. Inspired by many fields in computational science that have attained commanding results through deep learning techniques (DL), this paper envisions a methodology to apply DL techniques in MFA. We first conjecture that, in the future, CMA will attain more consistent and reliable results, thus replacing human analysis and stimulating the development of related fields. To achieve this higher goal, this paper stipulates two principles that need to be followed by subfields' results in CMA, namely, (i) conforming to the human cognition of music, and (ii) being reusable in other research investigations. These principles ensure that the results are trustworthy and reusable, which serve as the foundation for future exploration in other fields related to music.

Based on these two principles, we conduct a preliminary verification of our vision, proposing a BERT-based model (MTBert) for the *fugue* form's symbolic files (sequence processing). The results show that although the DL-based method still has some limitations, it can achieve better performance than the non-DL approach, which is an excitation for the future application of DL methods in MFA. Meanwhile, we also suggest combining the feature of the key transition in MF with DL for further complex MFA.

Finally, we systematically identify and address the challenges of MFA, and conclude them into two broad categories: music and non-music. The music category refers to the challenges caused by the particularity and complexity of MF as well as the features used for processing different music carriers. In contracts, the non-music category refers to two challenges of using DL on the complex MFA task, namely, the vague definition and complexity of the music form. When these challenges are addressed, the development of many fields related to music form can be further promoted, such as music education, music composition, music emotion/genre classification, making chord and downbeat estimation, psychology, Covid-19 treatment, heart disease treatment, etc.

References

1. Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H.: State-of-the-art in artificial neural network applications: a survey. Heliyon **4**(11), e00938 (2018)
2. Allegraud, P., et al.: Learning sonata form structure on mozart's string quartets. Trans. Int. Society Music Inform. Retrieval (TISMIR) **2**(1), 82–96 (2019)
3. Anagnostopoulou, C., Buteau, C.: Can computational music analysis be both musical and computational? J. Math. Music **4**(2), 75–83 (2010)
4. Arnold, J.M.: The role of chromaticism in Chopin's sonata forms: a Schenkerian view. Northwestern University (1992)
5. Basiri, M.E., Nemati, S., Abdar, M., Cambria, E., Acharya, U.R.: Abcdm: an attention-based bidirectional cnn-rnn deep model for sentiment analysis. Futur. Gener. Comput. Syst. **115**, 279–294 (2021)

6. Bergstrom, T., Karahalios, K., Hart, J.C.: Isochords: visualizing structure in music. In: Proceedings of Graphics Interface 2007, pp. 297–304 (2007)
7. Bigo, L., Giraud, M., Groult, R., Guiomard-Kagan, N., Levé, F.: Sketching sonata form structure in selected classical string quartets. In: ISMIR 2017-International Society for Music Information Retrieval Conference (2017)
8. Buccoli, M., Zanoni, M., Sarti, A., Tubaro, S., Andreoletti, D.: Unsupervised feature learning for music structural analysis. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 993–997. IEEE (2016)
9. Buisson, M., Mcfee, B., Essid, S., Crayencour, H.C.: Learning multi-level representations for hierarchical music structure analysis. In: International Society for Music Information Retrieval (ISMIR) (2022)
10. Burgoyne, J.A., Fujinaga, I., Downie, J.S.: Music information retrieval. A new companion to digital humanities, pp. 213–228 (2015)
11. Carr, C., Odell-Miller, H., Priebe, S.: A systematic review of music therapy practice and outcomes with acute adult psychiatric in-patients. PLoS ONE **8**(8), e70252 (2013)
12. Chan, W.Y., Qu, H., Mak, W.H.: Visualizing the semantic structure in classical music works. IEEE Trans. Visual Comput. Graphics **16**(1), 161–173 (2009)
13. Chawin, D., Rom, U.B.: Sliding-window pitch-class histograms as a means of modeling musical form. Trans. Int. Society for Music Inform. Retrieval **4**(1), (2021)
14. Chen, P., Zhao, L., Xin, Z., Qiang, Y., Zhang, M., Li, T.: A scheme of midi music emotion classification based on fuzzy theme extraction and neural network. In: 2016 12th International Conference on Computational Intelligence and Security (CIS), pp. 323–326. IEEE (2016)
15. Cheng, T., Smith, J.B., Goto, M.: Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 106–110. IEEE (2018)
16. Chew, E.: Cosmos: Computational shaping and modeling of musical structures. Front. Psychol. **13** (2022). https://doi.org/10.3389/fpsyg.2022.527539
17. Chillara, S., Kavitha, A., Neginhal, S.A., Haldia, S., Vidyullatha, K.: Music genre classification using machine learning algorithms: a comparison. Int. Res. J. Eng. Technol. **6**(5), 851–858 (2019)
18. Clercq, T.d.: Embracing ambiguity in the analysis of form in pop/rock music, 1982–1991. Music Theory Online **23**(3), (2017)
19. Corazza, G.E., Agnoli, S., Martello, S.: Counterpoint as a principle of creativity: extracting divergent modifiers from'the art of fugue'by johann sebastian bach. Musica Docta **4**, 93–105 (2014)
20. Dai, S., Jin, Z., Gomes, C., Dannenberg, R.B.: Controllable deep melody generation via hierarchical music structure representation. arXiv preprint arXiv:2109.00663 (2021)
21. De Prisco, R., et al: Music plagiarism at a glance: metrics of similarity and visualizations. In: 2017 21st International Conference Information Visualisation (IV), pp. 410–415. IEEE (2017)
22. De Prisco, R., Malandrino, D., Pirozzi, D., Zaccagnino, G., Zaccagnino, R.: Understanding the structure of musical compositions: is visualization an effective approach? Inf. Vis. **16**(2), 139–152 (2017)
23. Dent, E.J.: Binary and ternary form. Music Lett. **17**(4), 309–321 (1936)
24. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

25. Devlin, K., Alshaikh, J.T., Pantelyat, A.: Music therapy and music-based interventions for movement disorders. Curr. Neurol. Neurosci. Rep. **19**, 1–13 (2019)
26. Dirst, M., Weigend, A.S.: On completing js bach's last fugue. Time Series Prediction: Forecasting the Future and Understanding the Past, pp. 151–177 (1994)
27. Fuentes, M., McFee, B., Crayencour, H.C., Essid, S., Bello, J.P.: A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 481–485. IEEE (2019)
28. Giraud, M., Groult, R., Leguy, E., Levé, F.: Computational fugue analysis. Comput. Music. J. **39**(2), 77–96 (2015)
29. Giraud, M., Groult, R., Levé, F.: Subject and counter-subject detection for analysis of the well-tempered clavier fugues. In: Aramaki, M., Barthet, M., Kronland-Martinet, R., Ystad, S. (eds.) CMMR 2012. LNCS, vol. 7900, pp. 422–438. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41248-6_24
30. Hernandez-Olivan, C., Beltran, J.R., Diaz-Guerra, D.: Music boundary detection using convolutional neural networks: a comparative analysis of combined input features. Int. J. Interact. Multimedia Artif. Intell. **7**(2), 78 (2021). https://doi.org/10.9781/ijimai.2021.10.005
31. Huang, C.Z.A., Cooijmans, T., Roberts, A., Courville, A., Eck, D.: Counterpoint by convolution. arXiv preprint arXiv:1903.07227 (2019)
32. Jackendoff, R., Lerdahl, F.: The capacity for music: what is it, and what's special about it? Cognition **100**(1), 33–72 (2006)
33. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of The IEEE Conference On Computer Vision And Pattern Recognition, pp. 5308–5317 (2016)
34. Jin, C., Tie, Y., Bai, Y., Lv, X., Liu, S.: A style-specific music composition neural network. Neural Process. Lett. **52**, 1893–1912 (2020)
35. Jun, S., Hwang, E.: Music segmentation and summarization based on self-similarity matrix. In: Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, p. 4. No. 82 in ICUIMC '13, Association for Computing Machinery, New York, NY, USA (2013)
36. Kao, W.T., Lee, H.Y.: Is bert a cross-disciplinary knowledge learner? a surprising finding of pre-trained models' transferability. arXiv preprint arXiv:2103.07162 (2021)
37. Kenner, J., Baker, F.A., Treloyn, S.: Perspectives on musical competence for people with borderline personality disorder in group music therapy. Nord. J. Music. Ther. **29**(3), 271–287 (2020)
38. Kumar, C., Dutta, S., Chakborty, S.: Musical cryptography using genetic algorithm. In: 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014], pp. 1742–1747. IEEE (2014)
39. Lawes, M.: Creating a covid-19 guided imagery and music (gim) self-help resource for those with mild to moderate symptoms of the disease. Approaches: An Interdisciplinary Journal of Music Therapy, pp. 1–17 (2020)
40. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
41. Lewin, D.: Notes on the opening of the f# minor fugue from wtci. J. Music Theor. **42**(2), 235–239 (1998)
42. Manning, C.D.: Computational linguistics and deep learning. Comput. Linguist. **41**(4), 701–707 (2015)

43. Marandi, Y.M.H., Sajedi, H., Pirasteh, S.: A novel method to musicalize shape and visualize music and a novel technique in music cryptography. Multimedia Tools Appl. **80**, 7451–7477 (2021)
44. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. MIT press (2010)
45. Marsden, Alan: Music analysis by computer: ontology and epistemology. In: Computational Music Analysis, pp. 3–28. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25931-4_1
46. Mauch, M., Levy, M.: Structural change on multiple time scales as a correlate of musical complexity, pp. 489–494 (01 2011)
47. Meredith, D.: Music analysis and point-set compression. J. New Music Res. **44**(3), 245–270 (2015)
48. Miller, R.I.M.: Unity and contrast: A study of Ludwig van Beethoven's use of variation form in his symphonies, string quartets and piano sonatas. University of Glasgow (United Kingdom) (2003)
49. Müller, M.: Music Structure Analysis, pp. 167–236. Springer International Publishing, Cham (2015)
50. North, A.C., Hargreaves, D.J., Hargreaves, J.J.: Uses of music in everyday life. Music. Percept. **22**(1), 41–77 (2004)
51. Panda, R., Malheiro, R.M., Paiva, R.P.: Audio features for music emotion recognition: a survey. IEEE Trans. Affective Comput, 99, 1–1 (2020)
52. Pang, T.H.: The variation technique in selected piano works of Haydn, Mozart, Beethoven and Schubert: A performance project. University of Maryland, College Park (1998)
53. Paulus, J., Müller, M., Klapuri, A.: Audio-based music structure analysis. In: Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, pp. 625–636 (01 2010)
54. Pereira, R.M., Costa, Y.M., Aguiar, R.L., Britto, A.S., Oliveira, L.E., Silla, C.N.: Representation learning vs. handcrafted features for music genre classification. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
55. Pipalia, K., Bhadja, R., Shukla, M.: Comparative analysis of different transformer based architectures used in sentiment analysis. In: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 411–415. IEEE (2020)
56. Prout, E.: Fugue. Library Reprints (1891)
57. Ratner, L.: Harmonic aspects of classic form. J. Am. Musicol. Soc. **2**(3), 159–168 (1949)
58. Roy, S., Biswas, M., De, D.: imusic: a session-sensitive clustered classical music recommender system using contextual representation learning. Multimedia Tools Appl. **79**, 24119–24155 (2020)
59. Sheldon, D.A.: The stretto principle: some thoughts on fugue as form. J. Musicol. **8**(4), 553–568 (1990)
60. Shi, E.R., Zhang, Q.: A domain-general perspective on the role of the basal ganglia in language and music: Benefits of music therapy for the treatment of aphasia. Brain Lang. **206**, 104811 (2020)
61. Sutton, E.: Virginia Woolf and Classical Music: Politics, Aesthetics. Edinburgh University Press, Form (2013)
62. Tavares, J.M.R., Jorge, R.M.N., et al.: Topics in Medical Image Processing and Computational Vision. Springer (2013). https://doi.org/10.1007/978-94-007-0726-9

63. Umer, S., Mondal, R., Pandey, H.M., Rout, R.K.: Deep features based convolutional neural network model for text and non-text region segmentation from document images. Appl. Soft Comput. **113**, 107917 (2021)
64. Verma, P.K., Agrawal, P., Madaan, V., Prodan, R.: Mcred: multi-modal message credibility for fake news detection using bert and cnn. Journal of Ambient Intelligence and Humanized Computing, pp. 1–13 (2022). DOI: https://doi.org/10.1007/s12652-022-04338-2
65. Wang, W., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. arXiv preprint arXiv:2211.05778 (2022)
66. Webster, J.: Schubert's sonata form and brahms's first maturity. Nineteenth-Century Music, pp. 18–35 (1978)
67. Wen, R., Chen, K., Xu, K., Zhang, Y., Wu, J.: Music main melody extraction by an interval pattern recognition algorithm. In: 2019 Chinese Control Conference (CCC), pp. 7728–7733. IEEE (2019)
68. Wu, J., Liu, X., Hu, X., Zhu, J.: Popmnet: generating structured pop music melodies using neural networks. Artif. Intell. **286**, 103303 (2020)
69. Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional BERT contextual augmentation. In: Rodrigues, J.M.F., et al. (eds.) ICCS 2019. LNCS, vol. 11539, pp. 84–95. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22747-0_7
70. Young, J.O.: How classical music is better than popular music. Philosophy **91**(4), 523–540 (2016)
71. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015–1022. IEEE (2019)