PQG-A2SA: Performance Quantification Guided Audio-to-Score Alignment for Orchestral Music

Zhicheng Lian, Haonan Cheng, *Member, IEEE*, and Jiawan Zhang, *Senior Member, IEEE*

***Abstract*—Audio-to-score alignment is a multi-modal task that aims at generating an accurate mapping between symbolic and signal-level representations of musical signals, which is important for music performance analysis and retrieval. Among numerous music genres, orchestral music is a category of music with complex performance characteristics such as multi-instrument, non-percussive instrument and music expressiveness. However, previous methods do not take sufficient account of the performance characteristics of orchestral music, leading to limitations in alignment accuracy on orchestral music of these methods. To solve this problem, we present a performance quantification guided audio-to-score alignment (PQG-A2SA) method with high alignment accuracy for orchestral music at note-level. Specially, the PQG-A2SA contains two parts, namely an Inter Onset Interval (IOI) guided conditionally-constrained Dynamic Time Wrapping (DTW) and an articulation guided onset and offset detection. Different from the previous work, the IOI-guided conditionally-constrained DTW is designed to achieve a preliminary mapping between symbolic and chord-level representations of musical signals. In the second module, the onset and offset detection model under different musical articulations are established, thus refining the alignment results. We provide extensive experimental validation and analysis of our method. Our PQG-A2SA method can improve 9.0% in onset align rate and 17.5% in offset align rate at most compared with the state-of-the-art methods.**

***Index Terms*—Audio-to-score alignment, orchestral music, note-level, dynamic time wrapping, music information retrieval.**

I. Introduction

Audio-to-score alignment (A2SA) has been an active area of research in the Music Information Retrieval (MIR) community for decades, which can automatically find the optimal mapping between a performance and the corresponding symbolic score for a given piece of music. A2SA is widely used in tasks such as music analysis [1], [2], [3] and music source separation [4], [5], [6], etc. One of the key challenges in the field of A2SA is how to accurately detect note onsets and offsets from polyphonic mixtures. In this article, targeting the orchestral music with complex polyphony and music expressiveness, we propose a refined A2SA method based on sufficient quantitative information of performances.

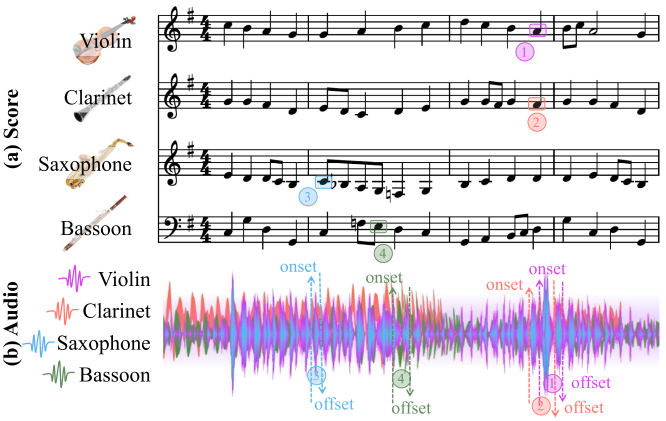


Fig. 1. Overview of orchestral music A2SA. (a) Illustrates the score for instruments: violin, clarinet, saxophone, and bassoon. (b) Presents the waveform mixtures of the four instruments in an ensemble. The numbers in circles indicate the correspondence between different notes and audio intervals. In our task, we aim to construct mappings between orchestral performance audio and the corresponding symbolic score.

Recent advances in the A2SA can be classified into two categories, namely online A2SA and offline A2SA. Online A2SA methods, also known as score following, allow for real-time timing estimation based on Hidden Markov Models (HMM) [7], online Dynamic Time Wrapping (DTW) [8], etc. However, methods in this category typically achieve lower alignment accuracies because only partial performance information can be used. Offline A2SA methods [9], [10] can take advantage of the full performance information, which generally enables a higher degree of accuracy. In order to deal with the asynchronies of notes in polyphonic music and further minimize the error between the score and audio, note-level A2SA methods [11], [12], [13] are proposed by adjusting the note onsets and offsets.

In note-level A2SA tasks, orchestral music alignment - the task of multi-instrumental polyphonic alignment (shown in Fig. 1), is particularly challenging due to the issues as follows:

- *Instrument issue:* The task is complicated in multi-instrumental signals that contain rich variations in timbre

Manuscript received 8 August 2022; revised 23 March 2023; accepted 2 May 2023. Date of publication 17 May 2023; date of current version 24 May 2023. This work was supported in part by the Natural Science Foundation of China under Grants 62172295 and 62201524, and in part by the Fundamental Research Funds for the Central Universities under Grant CUC22GZ002. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Chuang Gan. *(Zhicheng Lian and Haonan Cheng, contributed equally to this work.) (corresponding author: Jiawan Zhang.)*

Zhicheng Lian is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: lianzhicheng@tju.edu.cn).

Haonan Cheng is with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China (e-mail: haonancheng@cuc.edu.cn).

Jiawan Zhang is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China, and also with the Tianjin Cultural Heritage Conservation and Inheritance Engineering Technology Center and Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage, Tianjin University, Tianjin 300072, China (e-mail: jwzhang@tju.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3277290

from the employed orchestral instruments. Besides, for non-percussive instruments, onset and offset detection of orchestral instruments are difficult due to the gentle slope of the notes' attack and release phase.

- *Music expressiveness issue:* As a type of complex polyphonic music, orchestral music typically contains a variety of articulations and performance techniques that enhance the difficulty of A2SA.

Targeting the aforementioned issues, recent research advances refine orchestral alignment results by combining audio and image processing techniques [13], or by extracting the average harmonic structure of the instrument [14]. Although these works have made great progresses, there remains the limitation of alignment accuracy, especially for the offsets which cannot be handled effectively. Considering the performance characteristics of the orchestral music as an important source of information which may be not sufficiently utilized, we aim to accurately align the orchestral music by leveraging this information.

To this end, we propose a performance quantification guided audio-to-score alignment (PQG-A2SA) method which can effectively improve the alignment accuracy of orchestral music at note-level. PQG-A2SA framework consists of two stages. The first stage aims to generate a preliminary mapping between symbolic and chord-level representations of musical signals for further alignment. Specifically, a conditionally-constrained DTW is designed based on Inter Onset Interval (IOI) information guided duration constraint. In the second stage, a musical articulation guided onset and offset detection method is proposed for refining the alignment accuracy. The detection method is based on various musical articulations by establishing quantification schemes. Experiments on orchestral pieces show that the proposed PQG-A2SA method significantly outperforms state-of-the-art methods. In summary, our technical contributions are as follows:

- A two-stage method to provide high accurate note-level alignment is proposed by exploiting the performance quantification as guidance for A2SA of orchestral music.
- An IOI-guided conditionally-constrained DTW algorithm is proposed which can leverage the local and global information of audio recording and symbolic score sufficiently.
- We formalize temporal expressions and musical articulations of the orchestral music to bridge the gap between performance experience and quantitative method.

II. Related Work

In this section, we briefly review the recent advances of audio-to-score alignment, as well as the related works for characteristics of orchestral performance audio and scores.

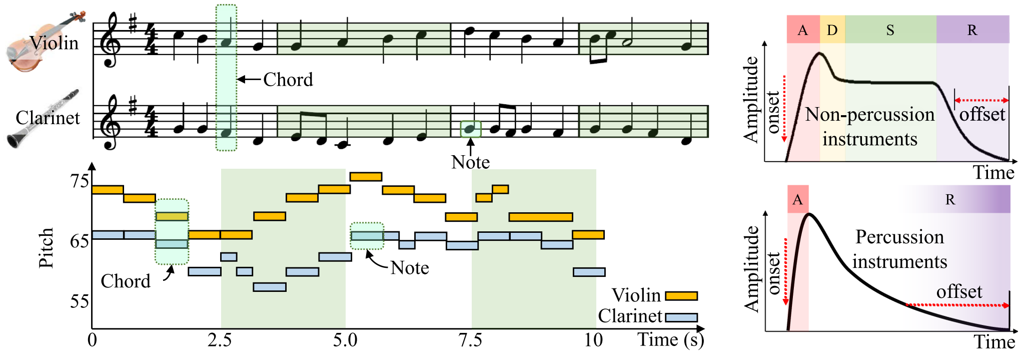
A. Audio-to-Score Alignment

A2SA is a multi-modal task which aims at constructing map between the performance audio and the corresponding symbolic score. The A2SA task is very challenging which includes many issues such as real-time capability of the algorithm [7], [8], identification of structural differences [15], robustness [16], [17], and accuracy of alignment, etc. In this work, we focus on the accuracy issue and discuss advances in this branch.

In earlier works, Orio and Schwarz [18] propose a DTW based alignment method, by utilizing spectral peak structure for local distance computation. Subsequently, DTW based researches are developed from different features, such as chroma representation [19], attack plus sustain note modeling [20], a combination of chroma-based features and onset-based features [9], acoustical features [21], etc. Apart from these methods, Joder et al. [22] introduce the use of Conditional Random Fields (CRFs) for the A2SA task. This method allows for the use of more flexible observation functions. Later, Joder et al. [23], [24] further propose an improved mapping method for A2SA. Recently, researchers enable adaptable A2SA for different playing environments via particle filter [25], observation model [26], Siamese networks [10], etc., while some researches focus on specific performing environments or music styles [27], [28], [29]. The above studies improve the accuracy of alignment by working on the level of time frames. As a result, these methods cannot identify the asynchronies between musical events that are notated as simultaneities in the score.

To further address the asynchrony problem, researches focus on methods of A2SA at note-level, that is, identify the onsets and offsets of all of the notes in score-notated simultaneities. Niedermayer [30] proposes a two-stage alignment method based on DTW and non-negative matrix factorization (NMF) for piano performance music, in which coarse alignment is performed by DTW and then note onsets are locally estimated by NMF. Later, Niedermayer and Widmer [31] further improve the accuracy of alignment by introducing anchor notes. Wang et al. [32] propose a pitch-by-time format called piano-roll feature for note-level A2SA. Devaney [11] introduces a multi-pass A2SA method which effectively improves the accuracy of alignment for monaural polyphonic recordings. Wang et al. [33] propose a method that handles asynchronies between the melody and the accompaniment by a multi-dimensional variant of DTW. However, the aforementioned methods mainly target single-instrument polyphonic music and ignore the case of multi-instrument polyphonic music.

For multi-instrumental polyphonic music, diversity of performance styles and timbral characteristics among different instruments makes it far more difficult for the A2SA task. In order to address the multi-instrument issue, Miron et al. [13] design a refined A2SA method based on combined audio and image processing techniques for orchestral music at note-level. This method gives a solution for predicting offsets of the notes, but the accuracy of the method is limited for alignment of offsets. Wang et al. [14] improve the accuracy of alignment for orchestral note onsets, benefiting from the computation of average harmonic structure learned from each source. But this method can not solve the problem of note offset. Simonetta et al. [12] present an automatic music transcription based deep learning method for A2SA. However, this method is not reliable enough for orchestral music. As a result, [13] and [14] still maintain the state-of-the-art performance for orchestral note offset and note onset alignment.



(a) Score and MIDI.

(b) The ADSR envelops.

Fig. 2. Illustration of orchestral music anatomy. (a) Presents the scores and corresponding MIDI representations of two instruments. (b) Shows the differences in the ADSR envelopes of percussion and non-percussion instruments.

Instead of audio processing paradigm widely used in prior methods, our proposed PQG-A2SA method takes the characteristics of orchestral performance into consideration by formalizing temporal expressions and performance techniques. With this design, the PQG-A2SA method can effectively refine the alignment of onsets and offsets in orchestral recordings.

B. Orchestral Music Anatomy

Music representation: Most previous works on A2SA take the score represented in Musical Instrument Digital Interface (MIDI). A MIDI file is a semantic encoding of musical signals (as shown in Fig. 2(a)) by specifying basic note parameters including time, pitch, volume and duration, together with parameters controlling instrument type, sound effect, track information, as well as music meta data [34]. In music notation, a note is a symbol denoting a musical sound with its pitch and duration, while a chord is a harmonic set of pitches consisting of multiple notes that are heard as if sounding simultaneously.

Onset and offset: One of the key challenges in orchestral A2SA is the accurate detection of note onset and offset. Onset and offset represent the beginning and ending of a musical note. In audio signals, attack-decay-sustain-release (ADSR) model can describe the energy envelope of a musical note in the performance (shown in Fig. 2(b)). As non-percussion instruments, orchestral instruments usually have soft onsets, compared with percussion instruments with hard onsets like the piano [35]. Since a soft onset generally has a gentle attack (top row in Fig. 2(b)), unlike the hard onset with sharp energy rise, the peak point in ADSR envelope cannot be directly approximated as onset, which increases the difficulty of accurate detection. In the case of offset, computationally estimation for notes of non-percussive instruments is also difficult, due to the the similar shapes of the decay and the release parts. Moreover, it is more complicated in real-world signals that contain rich variations in the orchestral instruments and musical articulations, which would shape the ADSR envelope in different ways [36]. Previous methods suffer from insufficient accuracy in terms of onset and offset in orchestral music, thus we explore and improve upon this limitation.

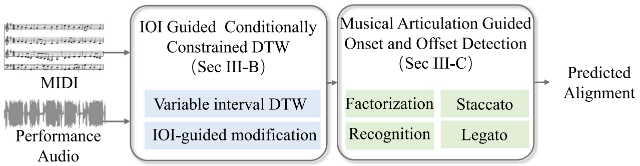


Fig. 3. An overview of our proposed PQG-A2SA method.

Musical articulation: Musical articulation is a fundamental musical conception that determines how notes are performed. It plays an important role in the performance characteristics to make the music expressive [37]. Musical articulations primarily structure an note's start and end, determining the length of its sound and the shape of its attack and decay. There are many types of musical articulation including staccato, legato, non legato, tenuto, marcato, etc. Each of them has a different effect on how the note is played. The most commonly used articulations are staccato and legato. Staccato refers to a note played shortened or detached. Legato refers to notes played smoothly and connected. In this work, we formalize typical musical articulations for onset and offset detection.

III. PQG-A2SA for Orchestral Music

In this section, we first present the formal formulation of the A2SA problem as an optimisation problem for orchestral music with multi-instruments. Then, we illustrate how to address the optimisation problem based on quantifying music performance. The overall architecture of the proposed PQG-A2SA approach is illustrated in Fig. 3. The core contributions, Inter Onset Interval (IOI) guided conditionally-constrained DTW method and musical articulation guided onset and offset detection, are then elaborated in Section III-B and Section III-C, respectively, which effectively leverage the local and global information and bridge the gap between performance experience and quantitative method.

A. Problem Statement

Suppose that we are given a music score X and the corresponding orchestral music audio Y with I instruments. The

A2SA task can be denoted as constructing a mapping function $\tau(\cdot)$ which contains I sub-functions $\tau_i(\cdot)$. In particular, for the i-th instrument, the mapping function is

$$\tau_i : \{l_{n_{i,1}}, \ldots, l_{n_{i,j}}, \ldots, l_{n_{i,J_i}}\} \rightarrow [0, T), \tag{1}$$

where $l_{n_{i,j}}$ represents the onset and offset location of note $n_{i,j}$ and T is the time length of the performance music. $n_{i,j}$ represents the j-th note for the i-th instrument in the note sequences $\{\{n_{i,j}\}_{j=1}^{J_i}\}_{i=1}^{I}$ of the score. I is the number of the instruments and J_i is the number of notes for the i-th instrument.

In our task, we aim to construct the mapping $\tau(\cdot)$, so that differences between the mapped location of note onset time $\tau_i(l^{on}_{n_{i,j}})$ and the real onset time $t^{on}_{n_{i,j}}$ in the performance audio are as small as possible, and similar differences are all for offset times. Consequently, the A2SA task is transformed into an optimisation problem, namely finding the optimal mapping $\tau(\cdot)$ to minimise the error between $\tau_i(l^{on}_{n_{i,j}})$ and $t^{on}_{n_{i,j}}$, and between $\tau_i(l^{off}_{n_{i,j}})$ and $t^{off}_{n_{i,j}}$. In this view, we define the error of the onsets NE as follows:

$$NE = \sum_{i=1}^{I}\sum_{j=1}^{J_i} |t^{on}_{n_{i,j}} - \tau_i(l^{on}_{n_{i,j}})|, \tag{2}$$

and the error of offsets FE:

$$FE = \sum_{i=1}^{I}\sum_{j=1}^{J_i} |t^{off}_{n_{i,j}} - \tau_i(l^{off}_{n_{i,j}})|. \tag{3}$$

According to this definition, our final goal is to construct the optimal $\tau(\cdot)$ to minimize the A2SA error E_{A2SA} that consists of onset error NE and offset error FE:

$$E_{A2SA} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (|t^{on}_{n_{i,j}} - \tau_i(l^{on}_{n_{i,j}})| + |t^{off}_{n_{i,j}} - \tau_i(l^{off}_{n_{i,j}})|). \tag{4}$$

B. IOI-Guided Conditionally-Constrained DTW

In this section, we first construct a chord-level alignment $\hat{\tau}(\cdot)$ with onsets of chords aligned. This is due to the fact that accurate chord-level alignment may critically affect the estimation of the approximate time range of each performed note in the audio. Specifically, we propose an IOI-guided conditionally-constrained DTW approach which consists of two procedures: variable interval DTW and IOI-guided modification.

The core idea of our proposed IOI-guided conditionally-constrained DTW approach is to consider note duration information of score as a conditional weak constraint. Due to the general tempo deviations between score and real performance, performer's freedom to lengthen or shorten on notes within tolerable limits in the alignment process may not be uniformly constrained by the note duration information in the score. But previous DTW approaches for A2SA often impose uniform tempo constraints by rewarding diagonal movement or penalizing non-diagonal movement in the whole alignment path. The uniform duration constraint can be meaningful for a robust alignment compared with the one without it, but may be not that useful for improving accuracy. So unlike previous DTW methods, our designed conditionally-constrained DTW contains a tempo-free and a duration-constrained DTW procedure, by which the duration constraints from score are conditionally imposed on segments with unexpected tempo deviations. In this way, we can get higher accuracy in the alignment with robustness maintained. Specifically, as shown in Fig. 4, the note duration in the score are temporally ignored to achieve a tempo-free DTW alignment with variable interval in the first procedure. Then in the second procedure, the quantified tempo information of performance is extracted from the alignment. Based on the IOI analysis, segments with unexpected tempo deviations are realigned with the duration-constrained DTW.

Variable Interval DTW: Firstly, the chroma feature of the score and audio are extracted. For the score, we use the chord-level representation where the chroma feature of chord sequence is extracted without duration information. Specifically, the input score X can be decomposed into a chord sequence $\{c_i\}_{i=1}^{L_{chord}}$, where L_{chord} represents the number of the chords. Then we transform the pitch information of chords into a chroma feature sequence $\{c_i^{crm}\}_{i=1}^{L_{chord}}$, where $c_i^{crm} \in \mathbb{R}_+^M$ and $M = 12$ is the number of pitch classes. The 12-dimensional chroma vector represents the semitones in an octave which is obtained by stacking pitches of notes in the chord on corresponding pitch class with L2-normalization.

For the audio, the chroma feature with variable interval is computed in two steps. In the first step, chroma feature of audio frames are extracted. The leading and trailing silence from the audio Y is trimmed to concentrate on the full music part of audio. The trimmed audio is computed in frames $\{f_i\}_{i=1}^{L_{frame}}$, where L_{frame} is the number of total frames. And then a Constant Q Transform (CQT) based chroma feature $\{f_i^{crm}\}_{i=1}^{L_{frame}}$ is extracted for each audio frame. In the second step, a temporally-constrained agglomerative clustering routine is adopted to partition audio frames into $L_{cluster}$ contiguous clusters $\{cl_i\}_{i=1}^{L_{cluster}}$. The number of clusters is computed according to the length of chords $L_{cluster} = \lfloor k_{cl} \cdot L_{chord} \rfloor$ where $k_{cl} \in \mathbb{R}$ and $k_{cl} \geq 1$ is a variable coefficient that satisfies $L_{chord} \leq L_{cluster} \leq L_{frame}$. Then the average chroma feature $\{cl_i^{crm}\}_{i=1}^{L_{cluster}}$ is computed for the sequence of clusters. In this procedure, the length of audio for DTW is reduced from L_{frame} frames to $L_{cluster}$ clusters in order to concentrate chords onsets on key frames and reduce the computation cost for DTW. Specifically, because the onset of a chord often shows drastic changes on chroma features, key frames with drastic chroma changes can be highlighted by temporal clustering as the boundaries of the clusters. While only key frames are set to be candidates for chords' onsets, the alignment can be easier to focus on these frames with the cost for computing other frames saved.

Based on the chord sequence for score and the cluster sequence for audio, we further compute the DTW-based alignment. The chroma feature sequences $\{c_i^{crm}\}_{i=1}^{L_{chord}}$ for chords and $\{cl_i^{crm}\}_{i=1}^{L_{cluster}}$ for audio are taken as input. The accumulated cost matrix D between the sequences is computed as follows:

$$D(x, y) = d(x, y) + \min \begin{cases} D(x, y-1) \\ D(x-1, y-1) \end{cases}, \tag{5}$$

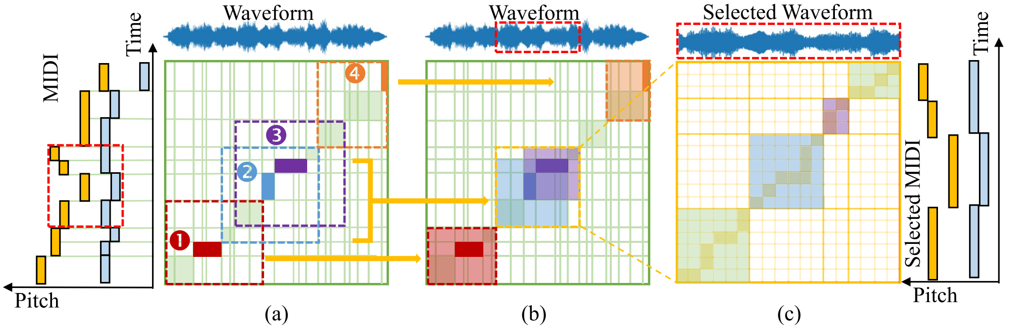


Fig. 4. Illustration of IOI-guided conditionally-constrained DTW. (a) Shows the process of variable interval DTW, where the input audio and MIDI are aligned by the DTW without duration constraints (the division of variable interval is indicated by green lines). Then, there are four unexpectedly deviated chords detected and numbered from 1 to 4. The dotted box with the same color indicates the local range of IOI-guided tempo check for each chord. (b) Illustrates the segments that need to be realigned. The 2nd (blue) and 3rd (purple) deviated chords are next to each other, so the segments are combined into a new one to be realigned. (c) Presents the duration-constrained DTW process of IOI-guided modification (the division of fixed interval is indicated by yellow lines). The 2nd and 3rd deviated chords are more accurately aligned after the modification.

where $x = 1, 2, \ldots, L_{chord}$ represents the index of chord sequence in the score and $y = 1, 2, \ldots, L_{cluster}$ represents index of cluster sequence in the audio. The function $d(\cdot)$ is a metric for computing euclidean distance between c_a^{crm} and cl_b^{crm}. Since we set the length of cluster sequence no less than the length of chords $L_{cluster} \geq L_{chord}$, only diagonal shifts and lateral shifts over the sequences are allowed in the alignment path. In this way, a chord will be aligned to distinct and continuous clusters based on the path backtracking on the accumulated cost matrix. The onset frame for each chord is obtained according to the boundary frame of the first cluster aligned to the chord. This alignment $\hat{\tau}^{(0)}(\cdot)$ is a preliminary onset frame estimation for chords in the score.

IOI-Guided Modification: In the procedure of variable interval DTW, some variations in the real performance may cause unexpected deviations in the alignment path without the note duration information. So in the procedure of IOI-guided modification, the aligned chords are conditionally selected and realigned with a duration-constrained DTW, where the condition is guided by the IOI-based tempo analysis of the performance.

In order to utilize the tempo information of the music performance, the alignment result is considered as a benchmark from which IOI is extracted to quantify the tempo information. Then we filter out the aligned segments that do not fit well into the performance regularity so as to realign them. We infer from the regularity of the performance that the tempo of the music performed locally does not produce large variations. Thus, methodologically, our idea is to quantify the tempo fluctuations by comparing the quantified performing velocity of the current chord with its nearby local velocity. Then the chords whose tempo changes drastically relative to the local tempo may deviate from the alignment and are selected.

Specifically, we first quantify the tempo information based on the aligned chord sequence $\{c_i\}_{i=1}^{L_{chord}}$ and the alignment result $\hat{\tau}^{(0)}(\cdot)$ of variable interval DTW:

$$v_{c_i} = \frac{\hat{\tau}^{(0)}(l_{c_{i+1}}^{on}) - \hat{\tau}^{(0)}(l_{c_i}^{on})}{l_{c_{i+1}}^{on} - l_{c_i}^{on}}, \tag{6}$$

Algorithm 1: IOI-Guided Conditionally-Constrained DTW.

```
Input:
    Score $X$ and audio $Y$;
Output:
Chord-level alignment $\hat{\tau}(\cdot)$;
Represent $X$ with chords $\{c_i\}_{i=1}^{L_{chord}}$ and segment $Y$ into frames $\{f_i\}_{i=1}^{L_{frame}}$;
Compute chroma feature on score chords $\{c_i^{crm}\}_{i=1}^{L_{chord}}$;
Compute chroma feature on audio frames $\{f_i^{crm}\}_{i=1}^{L_{frame}}$;
$\{cl_i\}_{i=1}^{L_{cluster}} \leftarrow TemporalCluster(\{f_i^{crm}\}_{i=1}^{L_{frame}})$;
Compute average chroma on audio clusters $\{cl_i^{crm}\}_{i=1}^{L_{cluster}}$;
update $\hat{\tau}^{(0)}$ by Eq.(5);
Initialize an empty set $U$ for unexpected chords;
for $i = 1$ to $L_{chord}$ do
    Compute $v_{c_i}$, $v'_{c_i}$, $r_{c_i}$;
    if $r_{c_i} < r_{low}$ or $r_{c_i} > r_{high}$ then
        $U \leftarrow U \cup \{c_i\}$;
    end if
end for
Combine continuous chords in $U$ as segments set $S$;
for each segment $s_i \in S$ do
    Compute chroma feature on score beats $\{b^{crm}\}_{s_i}$;
    Compute chroma feature on audio frames $\{f^{crm}\}_{s_i}$
    Update $\hat{\tau}$ by Eq.(9);
end for
return $\hat{\tau}(\cdot)$;
```

where v_{c_i} indicates the velocity for the chord c_i defined by the IOI of c_i in the audio relative to the score according to the alignment.

Then we quantify the local tempo as follows:

$$v'_{c_i} = \frac{\hat{\tau}^{(0)}(l_{c_{i+\Delta i}}^{on}) - \hat{\tau}^{(0)}(l_{c_{i-\Delta i}}^{on})}{l_{c_{i+\Delta i}}^{on} - l_{c_{i-\Delta i}}^{on}}, \tag{7}$$

where Δi is the local range of chords, and v'_{c_i} indicates the local velocity for chord c_i defined by the IOI of nearby continuous chords ranging from chord $c_{i-\Delta i}$ to $c_{i+\Delta i}$ in the audio relative to the score. We set $\Delta i = 4$ in the experiment. To show the velocity deviation of a chord from the local range, we utilize the velocity of the current chord relative to its local velocity to quantify the deviation of the performed chord based on the alignment:

$$r_{c_i} = \frac{v_{c_i}}{v'_{c_i}}. \tag{8}$$

Next, we determine whether there is a significant deviation from the current local velocity of the chord by the threshold r_{low} and r_{high}. Segments in the range $r_{low} \leq r_{c_i} \leq r_{high}$ are considered within tolerance, where $r_{low} \in (0, 1)$ and $r_{low} \cdot r_{high} = 1$. As shown in Fig. 4(a), segments that are out of range are treated as having unexpected deviations which are selected to be realigned. The chords are examined one by one for deviations, and the unexpectedly deviated chords are added into a set U.

Then we combine the continuous chords in U as segments to get a new set S as shown in Fig. 4(b). Since these segments may be misaligned due to the differences between the score and the music in the real performance, tempo information is locally introduced to obtain a more stable treatment. Specifically, for segment $s_i \in S$ containing chords from c_p to c_q, we realign the segment at frame-level in score ranging from $l^{on}_{c_{p-1}}$ to $l^{on}_{c_{q+1}}$ as x', and in audio ranging from $\hat{\tau}^{(0)}(l^{on}_{c_{p-1}})$ to $\hat{\tau}^{(0)}(l^{on}_{c_{q+1}})$ as y'. In the experiment, the score is segmented at 0.02 beat per interval so as to put in duration information for the alignment, and the audio is segmented to frames at a hop length of 23 ms.

Finally, the chroma feature in the range is accordingly computed for score beats $\{b^{crm}\}_{s_i}$ and audio frames $\{f^{crm}\}_{s_i}$. As shown in Fig. 4(c), we realign the score and audio sequences of each segment based on the DTW with chroma feature as input and Euclidean distance as metric for each $s_i \in S$:

$$D(x', y') = d(x', y') + \min \begin{cases} D(x', y' - 1) \\ D(x' - 1, y') \\ D(x' - 1, y' - 1) \end{cases}. \tag{9}$$

Moreover, the unexpectedly deviated segments are modified with backtracking on the alignment path. With the encouraged diagonal movements, a duration-constrained alignment is imposed on these segments to achieve a robust result. Then the alignment result $\hat{\tau}^{(i)}$ is updated with each segment s_i realigned. As a result, we can get an alignment $\hat{\tau}(\cdot)$ at chord-level to estimate the onsets of chords.

C. Articulation-Guided Onset and Offset Detection

After the chord-level alignment, there still remain two main issues: (1) Soft onset and offset issue. The onsets and offsets of notes performed by orchestral non-percussive instruments are hard to accurately detect due to the gentle note attack and release phase compared with percussive instruments. (2) Articulation issue. Flexibly used articulations on orchestral instruments, especially non-percussive instruments, can remarkably affect how a note is performed, thus causing subtle deviations of onsets and offsets and leading to the challenge of accurate detection. Aiming at these issues, we detect onsets and offsets for each note in the music signal through a quantified note performance model and propose an articulation guided onset and offset detection method to achieve an accurate alignment at the note-level.

Score-Informed Spectrogram Factorization: In order to achieve an accurate note-level alignment, we firstly aim to decompose the input music signal Y into pitch-based and instrument-dependent representation. Traditional score-informed NMF method [4] can effectively decompose piano music. Targeting on orchestral music, we propose a modified NMF-based spectrogram factorization method for an efficient decomposition to get the activation energy on pitches of instruments.

Specifically, The goal of NMF-based procedure is to find non-negative matrices W and H as a decomposition of V such that $V \approx \hat{V} = W \cdot H$, where matrix V is a magnitude spectrogram of audio signal, matrix W is template vectors for picking up the structure of the pitch-dependent spectral vectors, and H is an activation matrix that encodes when and how strongly the respective vectors are active. V is initialized with logarithmically compressed magnitude spectrogram $log(1 + |STFT(Y)|^{\top})$. Then, score information is combined to constrain the process of factorization. To initialize W, pitch sets $\{P_i\}_{i=1}^{I}$ for I instruments are individually computed, and a zero vector B is used to catch background noise. Therefore, pitch-based harmony templates $P_1, P_2, \ldots, P_I$ and B are stacked to be W. Next, based on the alignment $\hat{\tau}(\cdot)$ in the first stage, we set the corresponding entries from score in H to 1 with an tolerance window for onset $tol_{on} = 150$ ms and for offset $tol_{off} = 200$ ms, while all remaining entries are set to zero. After 150 NMF iterations, we enforce that all columns of the final template matrix W are normalized with the max-normalization in order to make activation energy in H on different templates more comparable.

Activation-Based Articulation Recognition: After computing the activation matrix H, we could know when and how strongly the energy is on a specific pitch of an instrument. In this step, we aim to recognize articulations of performed notes based on the activation matrix H. Due to the common characteristics and similar influence exerted on the notes' performance for non-percussive instruments, articulations are used as a classification criteria to model each note in a more fine grained level. The articulations are then divided into two categories: namely staccato (including non legato) and legato for recognition.

We first preprocess the average energy information and the region of interest (ROI) for computation on a pair of continuous notes. With regard to notes $n_{i,j}$ and $n_{i,j+1}$ of instrument i, we roughly locate their positions of frames in the audio by the chord-level alignment. As shown in Fig. 5, $f_a = \hat{\tau}(l^{on}_{n_{i,j}})$, $f_b = \hat{\tau}(l^{on}_{n_{i,j+1}})$, and $f_c = \hat{\tau}(l^{on}_{n_{i,j+2}})$ are used to represent the approximate onset and offset frames of the notes. Then the average energy of the notes are estimated from the activation matrix H over the corresponding frames:

$$\begin{cases} e_\alpha = \frac{1}{N_\alpha} \sum_{i=f_a}^{f_b} H_{r_\alpha, i} \\ e_\beta = \frac{1}{N_\beta} \sum_{i=f_b}^{f_c} H_{r_\beta, i} \end{cases}, \tag{10}$$

where r_α represents the pitch of note $n_{i,j}$ and r_β represents the pitch of the note $n_{i,j+1}$. And N_α indicates the number

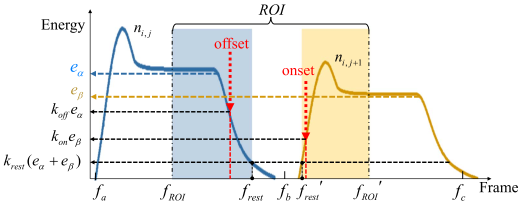


Fig. 5. Illustration of the staccato model. The energy envelop and average energy e_α, e_β of the two continuous notes $n_{i,j}$ and $n_{i,j+1}$ are individually shown in blue and yellow. Concentrating on ROI ranging from f_{ROI} to f_{ROI}', the rest segment in time range from f_{rest} to f_{rest}' is detected with upper limit of energy to $k_{rest}(e_\alpha+e_\beta)$. The offset f_{off} of note is then detected in the blue area, using the lower limit of energy to $k_{off}e_\alpha$. And the onset f_{on} of the following note is similarly detected in the yellow area.

of total frames that satisfy $H_{r_\alpha,f} > 0$ in the range of $[f_a, f_b]$. Accordingly, it is similar for N_β. To concentrate on the region for computation, we restrict the ROI for searching to an adaptive interval $[f_{ROI}, f_{ROI}']$, where $f_{ROI} = (f_a + f_b)/2$ and $f_{ROI}' = (f_b + f_c)/2$ are in the middle of each note. Then we compute the sum of the energy on each frame for the continuous notes to reflect the trend of total energy variations in the ROI that $Hs = H_{r_\alpha} + H_{r_\beta}$.

After computing the ROI frames and the average energy of notes, we then identify the note performance model as a staccato or legato by checking the existence of a continuous low-sum-energy segment on the notes' pitches. It is explained that staccato is manifested in the performance as the stopped blowing or bowing within a period of time, which may lead to a rest notated nowhere in the score. So we recognize the articulation by checking the length of the interval in which the sum of the energies on the pitches is less than the average energy of the two notes in a fixed proportion. Formally, we consider variable f_{rest}, f_{rest}' representing the beginning and ending of the rest segments between continuous notes, and parameter T_s representing the minimum length of the rest segment. The constrain is set that $f_{rest}, f_{rest}' \in [f_{ROI}, f_{ROI}']$ and $f_{rest}' - f_{rest} \geq T_s$, where the rest should not be too short in case of the possible fluctuation of energy. For each frame f in the range $f \in [f_{rest}, f_{rest}']$, if the sum of activation energy on the notes satisfy:

$$Hs_f < k_{rest}(e_\alpha + e_\beta), \tag{11}$$

where k_{rest} is the coefficient to constrain the maximum energy of the rest segment in proportion to the sum average energy of the notes, then we identify the note $n_{i,j}$ is performed as a staccato model (including non legato articulation). Otherwise, the note is performed as a legato model.

Staccato Model Computation: After recognizing the articulation of notes, onsets and offsets are predicted based on different articulation models. Since there are usually small variations of fluctuation on activation energy envelope in practice, it is inaccurate to directly use the energy greater than zero to indicate the onset or offset of the notes. So we combine the characteristics on different articulations and build a parametric model to accurately predict the onset and offset.

Algorithm 2: Articulation-Guided Onset and Offset Detection.

```
Input:
    Score $X$ and audio $Y$;
    Chord-level alignment $\hat{\tau}(\cdot)$;
Output:
    Note-level alignment $\tau(\cdot)$;
Represent $X$ with note sequences $\{\{n_{i,j}\}_{j=1}^{J_i}\}_{i=1}^{I}$;
$V \leftarrow log(1 + |STFT(Y)|^\top)$
Initialize pitch sets $\{P_i\}_{i=1}^{I}$ for $I$ instruments from $X$,
    and $B$ for background noise;
$W^{(0)} \leftarrow$
    $InitializePitchTemplate(P_1, P_2, \ldots, P_I, B)$
$H^{(0)} \leftarrow InitializeScoreActivation(X, \hat{\tau})$
$W, H \leftarrow NMF(V, W^{(0)}, H^{(0)})$
for $i = 1$ to $I$ do
    for $j = 0$ to $J$ do
        Compute $f_a, f_b, f_c, f_{ROI}, f_{ROI}', e_\alpha, e_\beta, Hs, Hd$;
        if exists $f_{rest}, f_{rest}'$ satisfy Eq.(11) then
            Compute $t_{off}, t_{on}$ by Staccato model;
        else
            Compute $t_{off}, t_{on}$ by Legato model;
        end if
        $\tau_i(l^{off}_{n_{i,j}}) \leftarrow t_{off}$, $\tau_i(l^{on}_{n_{i,j+1}}) \leftarrow t_{on}$;
    end for
end for
return $\tau(\cdot)$;
```

In the staccato model, we predict the offset and onset by using the activation energy of respective notes since the continuous notes are relatively independent with little or no overlap in the spectrogram. Specifically, to detect the offset of the note $n_{i,j}$, we find the frame preceding the frame f_{rest} where the energy of note is above a proportion of its average energy. As illustrated by the blue block in Fig. 5, we constrain the candidate frames $f \in [f_{ROI}, f_{rest}]$ that satisfy:

$$H_{r_\alpha,f} > k_{off} \cdot e_\alpha, \tag{12}$$

where k_{off} is a parameter to indicate the proportion of the average note energy for detecting the offset in a staccato model. Then we choose the closest frame to the beginning of rest segment:

$$f_{off} = argmin(|f_{rest} - f|). \tag{13}$$

Then the time t_{off} of frame f_{off} is predicted to be the offset.

Similarly, to detect the onset of the note $n_{i,j+1}$, we find the frame following the frame f_{rest}' where the energy of note is above a proportion of its average energy. As illustrated by the yellow block in Fig. 5, we constrain the candidate frames $f \in [f_{rest}', f_{ROI}']$ that satisfy:

$$H_{r_\beta,f} > k_{on} \cdot e_\beta, \tag{14}$$

where k_{on} is a parameter to indicate the proportion of the average note energy for detecting the onset. Then we choose the closest frame to the ending of rest segment:

$$f_{on} = argmin(|f - {f_{rest}}'|). \tag{15}$$

Then the corresponding time t_{on} of frame f_{on} is predicted to be the onset. It is noted that if the note follows or is followed by a segment of rest, including the beginning and ending note of the score, its onset or offset is predicted in a staccato model.

Legato Model Computation: The legato articulation on notes of non-percussive instruments is usually manifested as the change of fingering or bowing with notes played connected, which is expressed in the spectrogram as the overlapping of energy in a short period of time. Therefore, in the legato model, offset and onset are detected from energy relation between the continuous notes.

We first detect the onset of the note $n_{i,j+1}$, which is considered to be the alternation position between the continuous notes. So we use the the difference of energy on the continuous notes $Hd = H_{r_\alpha} - H_{r_\beta}$ to reflect the energy relation of the notes. Within the range $f \in [f_{ROI}, {f_{ROI}}']$, we constrain f by considering the difference of energy:

$$Hd_f \cdot Hd_{f+1} \leq 0, \tag{16}$$

where the energy of the following note $n_{i,j+1}$ exceed the preceding note $n_{i,j}$. Because there maybe some tiny energy fluctuations in the short time, the eligible f may not be unique. So we choose the one is closest to the chord-level aligned onset f_b to make the result more reliable:

$$f_{on} = argmin(|f_b - f|). \tag{17}$$

Then the time t_{on} of frame f_{on} is predicted to be the onset of note $n_{i,j+1}$.

The offset of the note $n_{i,j}$ is considered to be the position where the sum energy decreases within a short period of time before the onset. This can be explained as a momentary adjustment on fingering or bowing. Specifically, the parameter T_l is set to be a short period of time as a searching window for offset detection. Then we search for frames in the range $f \in [f_{on} - T_l, f_{on}]$ that satisfy:

$$Hs_f > Hs_{f+1}, \tag{18}$$

where the sum energy is decreasing as illustrated by green dashed box in Fig. 6. But also, the eligible f may not be unique. So we choose the farthest frame to the onset f_{on} which is considered to be the start position of the decreasing phase. The predicted offset frame of note $n_{i,j}$ is:

$$f_{off} = argmax(|f_{on} - f|). \tag{19}$$

Then the time t_{off} of frame f_{off} is predicted to be the offset. Additionally, for notes with the same pitch in legato, this model may not work well due to the fully overlapped envelope. So we use the chord-level alignment result t_b as the predicted offset t_{off} and onset t_{on} for this situation.

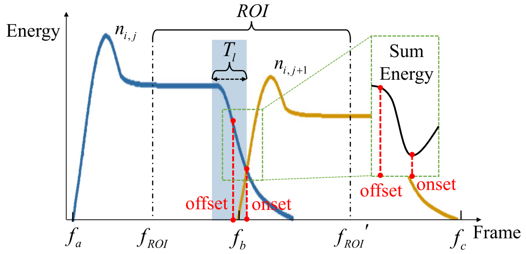


Fig. 6. Illustration of the legato model. The energy envelop of the two continuous notes $n_{i,j}$ and $n_{i,j+1}$ are shown in blue and yellow. The onset frame f_{on} of the following note is detected where the energy of the following note exceed the preceding note. The the offset frame f_{off} of the preceding note is detected in a short period of time T_l before f_{on}, where the sum energy of the two notes decreases, as shown in the green dashed box.

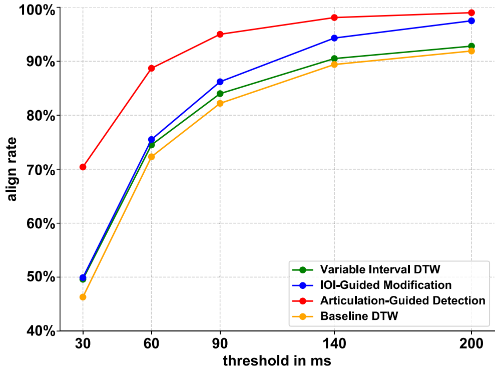


Fig. 7. Align rate of onset in different variants of the system and in comparison with a baseline DTW method.

TABLE I
VALUES OF PARAMETERS FOR THE SYSTEM

Parameter	Value	Description
k_{cl}	10.0	coefficient for cluster length
r_{low}	0.6	ratio for unexpected IOI deviation
k_{rest}	0.15	coefficient for rest energy
k_{off}	0.6	coefficient for offset energy
k_{on}	0.2	coefficient for onset energy
T_s	150 ms	shortest note off-onset interval of staccato
T_l	30 ms	longest note off-onset interval of legato

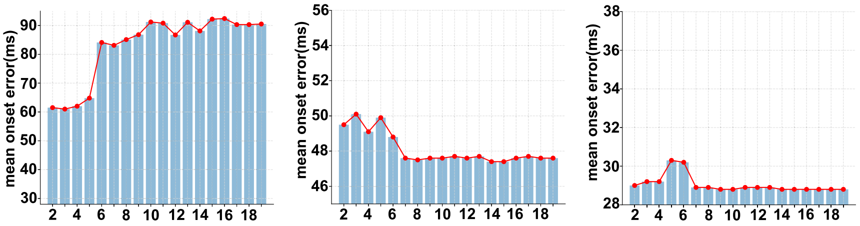
IV. EXPERIMENTS AND DISCUSSIONS

A. Experimental Setup

All our experiments are based on the orchestral music dataset Bach10 [6]. In the experiment, the audio is segmented to frames with hop size of about 23 ms. The values of parameters with the best performance are listed in Table I. The experiment is done on a computer of windows system with Intel(R) Core(TM) i7-10875H CPU and 16.0 GB RAM.

B. Model Ablation Studies

To validate the parameters of our model, we use align rate metric and mean error to qualify the performance of parameter choices. Considering the absolute time error e of each note, we use the mean value to measure the average proximity to the real



(a) VIDTW *MNE* of k_{cl} (b) IOIGM *MNE* of k_{cl} (c) AGOOD *MNE* of k_{cl}

Fig. 8. Illustration of mean onset error of variable parameter k_{cl} in different phases of the system. (a) Represents mean onset error of k_{cl} in VIDTW phase. (b) Represents mean onset error of k_{cl} in IOIGM phase. (c) Represents mean onset error of k_{cl} in AGOOD phase.

TABLE II
Comparison of Variants of Our Model

Method	Onset align rate ↑					Offset align rate ↑					Mean error (ms) ↓	
	30 ms	60 ms	90 ms	140 ms	200 ms	30 ms	60 ms	90 ms	140 ms	200 ms	onset	offset
Baseline	46.3%	72.3%	82.2%	89.4%	91.9%	-	-	-	-	-	85.0	-
VIDTW	49.6%	74.5%	84.0%	90.5%	92.8%	-	-	-	-	-	91.2	-
VIDTW+IOIGM	49.9%	75.5%	86.2%	94.3%	97.5%	-	-	-	-	-	47.6	-
Baseline+AGOOD	67.9%	83.8%	90.3%	93.9%	94.8%	64.2%	80.1%	86.0%	91.1%	93.9%	51.7	55.8
VIDTW+AGOOD	67.5%	84.4%	91.0%	94.5%	95.4%	63.1%	80.1%	85.8%	90.4%	92.4%	69.9	81.0
VIDTW+IOIGM+AGOOD	**70.4%**	**88.7%**	**95.0%**	**98.1%**	**99.0%**	**66.5%**	**83.8%**	**89.6%**	**94.6%**	**96.6%**	**28.8**	**39.1**

time position of onset:

$$\mathrm{MNE} = \frac{NE}{N_{note}}, \tag{20}$$

and offset:

$$\mathrm{MFE} = \frac{FE}{N_{note}}, \tag{21}$$

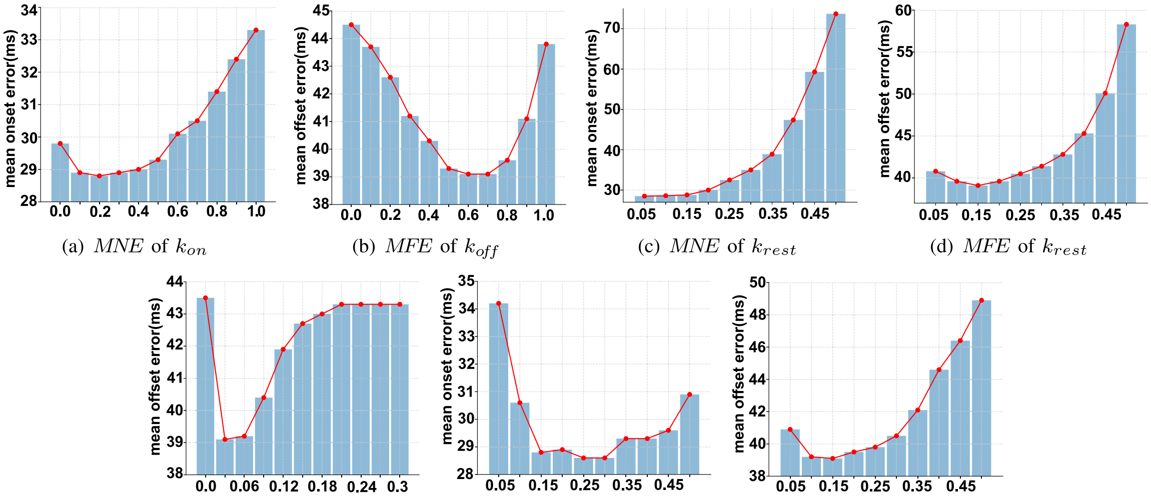
where N_{note} is the total number of notes in the score. The align rate is the portion of all events with e less than a pre-defined threshold, where we use the thresholds between 30 ms and 200 ms.

The ablation experiment is shown in Table II and Fig. 7, where VIDTW represents variable interval DTW, IOIGM represents IOI-guided modification and AGOOD represents articulation-guided onset and offset detection. By comparing the results of a baseline DTW method with the results of different variants, it is obvious to see that each part of our method has improved the alignment accuracy. Variable interval DTW is similar to the baseline DTW in terms of results, but can be significantly improved by IOI-guided modification, where the mean onset error is reduced from 91.2 ms to 47.6 ms. This provides a more accurate range of note positioning for subsequent procedures. Articulation-guided onset and offset detection works at the note-level, providing a significant accuracy improvement. The better results we achieve for alignment at the chord-level, the better final results will be obtained at the note-level. The best result reaches an onset error of 28.8 ms and an offset error of 39.1 ms, significantly exceeding the alignment accuracy that the baseline can provide. And it enables stable detection for most notes, with 95.1% of note onsets detected at an error threshold of 140 ms and 99% of note onsets detected at 200 ms. Even for note offset detection, which is considered a difficult problem, 94.6% and 96.6% of note offsets are detected at an error threshold of 140 ms and 200 ms, respectively.

In order to verify the reasonableness of the parameter selection, we design three experiments. We first validate the alignment effect at three different procedures when the parameter k_{cl} is in different value choices. As shown in Fig. 8, it is obvious to see that k_{cl} has a greater impact on the VIDTW procedure, and a choice of small value can improve the accuracy of the results. This is consistent with the scenario we envisioned, where the temporal clustering process concentrates the cluster boundaries on audio frames with drastic feature variations, thus making it easier to obtain a more accurate alignment. The result with different value choices is compensated to be closer by IOIGM and AGOOD. But too small value choice may lead to slight instability in the result, while a larger value will take more computation cost in the sequence alignment process. So the value 10 is chosen for the parameter as a balance.

Then we validate the effects of several adjustable parameters on the results in the AGOOD stage, respectively. As illustrated in Fig. 9, the optimal value of k_{rest} is 0.15, corresponding to the proportion of the notes' sum average energy for the rest segment. The optimal value of k_{off} is 0.6 to 0.7, corresponding to the note offset which is shown to be close to the release of note energy. And k_{on} is 0.2, corresponding to the note onset which is verified to be close to the start of the note energy rise. These are all consistent with the theoretical note model. For the minimum staccato rest interval T_s of 150 ms and the maximum legato interval T_l of 30 ms, the parameters are verified to be consistent with the intuitive note staccato and legato performance requirements.

Finally, we compare the effect in the IOIGM procedure with the change of parameter r_{low} and k_{cl}. It can be seen in Fig. 10 that even if we change the value of the parameter k_{cl}, we still get the optimal result with a stable value 0.6 of parameter r_{low}. The closer the value of the parameter r_{low} is to 1.0, the more segments with larger region are imposed duration constraint upon. So the choice of optimal value 0.6 illustrates that our proposed method



(a) *MNE* of k_{on} (b) *MFE* of k_{off} (c) *MNE* of k_{rest} (d) *MFE* of k_{rest}

(e) *MFE* of T_l (f) *MNE* of T_s (g) *MFE* of T_s

Fig. 9. Illustration of mean error with the change of different parameters in the phase of AGOOD. (a) Presents mean onset error of parameter k_{on}. (b) Presents mean offset error of parameter k_{off}. (c) Presents mean onset error of parameter k_{rest}. (d) Presents mean offset error of parameter k_{rest}. (e) Presents mean offset error of parameter T_l. (f) Presents mean onset error of parameter T_s. (g) Presents mean offset error of parameter T_s.

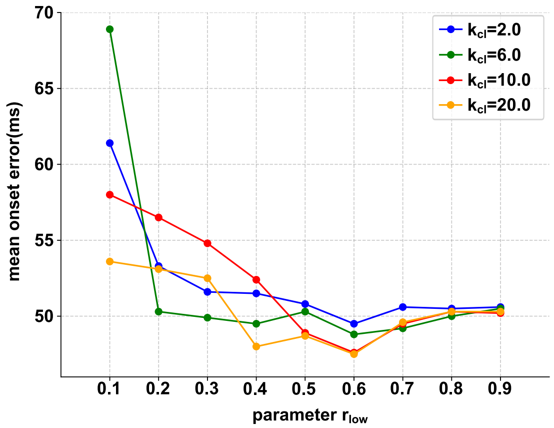


Fig. 10. Mean onset error of parameter r_{low} with changes of parameter k_{cl}.

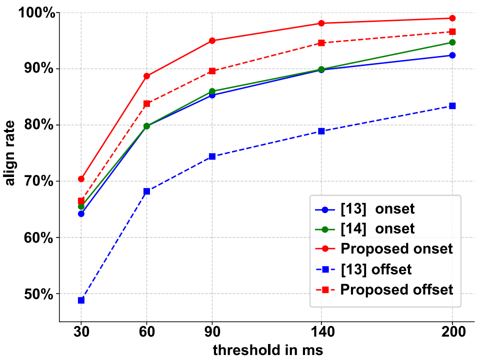


Fig. 11. Comparison with SOTA methods. The blue solid line, blue dashed line, green solid line, red solid line and red dashed line represent the onset alignment and offset alignment of Miron et al. [13], Wang et al. [14] and our method, respectively.

achieves the subtle balance between the local tempo variation and the overall duration constraint.

C. Comparison With the State-of-The-Arts

We compare the note-level alignment result of our proposed method with SOTA methods [13] and [14] based on Bach10 dataset. Among them, Miron's work [13] detects both onsets and offsets of notes, and Wang's work [14] detects only the onsets of notes. Fig. 11 shows the comparison results on the align rate. Our method outperforms the onset and offset align rate at error thresholds from 30 ms to 200 ms, achieving SOTA results. The align rate of onset is improved from about 65.5% to 70.4% at an error threshold of 30 ms, from about 86.0% to 95.0% at 90 ms, and from about 95.0% to 99.0% at 200 ms. The align rate of offset is significantly improved compared with the previous method. The align rate of offset improved from less than 50.0% to 66.5% at an error threshold of 30 ms, and from about 83.5% to 96.6% at 200 ms.

V. Conclusion and Future Work

In order to improve the accuracy of A2SA in the context of orchestral music, we propose a two-stage PQG-A2SA method. In the first stage, we first utilize the IOI to guide a conditionally-constrained DTW and achieve a preliminary mapping between symbolic and chord-level representations of musical signals. In the second stage, an articulation based onset and offset detection model under different musical articulations are established, thus significantly refining the alignment at note-level. As a result, we improve 9.0% in align rate of onset at an error threshold of 90 ms and improve 17.5% in align rate of offset at an error threshold of 30 ms compared with the SOTA method.

In this work, we find that our method may make some error in some notes with other performance techniques. For example, our method sometimes fail to accurately detect the offsets of notes performed using vibrato due to the wide fluctuation of energy envelope. In the future, we will introduce more performance technique and more quantified performance models to guide the accurate alignment of orchestral music.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] A. M. Wing, S. Endo, A. Bradbury, and D. Vorberg, "Optimal feedback correction in string quartet synchronization," *J. Roy. Soc. Interface.*, vol. 11, no. 93, 2014, Art. no. 20131125.

[2] R. Stables, S. Endo, and A. Wing, "Multi-player microtiming humanisation using a multivariate markov model," in *Proc. Conf. Digit. Audio Effects*, 2014, pp. 1–5.

[3] S. Wang, S. Ewert, and S. Dixon, "Identifying missing and extra notes in piano recordings using score-informed dictionary learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1877–1889, Oct. 2017.

[4] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2012, pp. 129–132.

[5] M. Miron, J. Janer, and E. Gómez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2017, pp. 55–62.

[6] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, Oct. 2011.

[7] T. Nakamura, E. Nakamura, and S. Sagayama, "Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 329–339, Feb. 2016.

[8] S. Dixon, "Live tracking of musical performances using on-line time wrapping," in *Proc. Conf. Digit. Audio Effects*, 2005, pp. 1–6.

[9] S. Ewert and M. Müller, "Refinement strategies for music synchronization," in *Proc. Int. Symp. Comput. Music Modelling Retrieval*, 2009, pp. 147–165.

[10] R. Agrawal and S. Dixon, "Learning frame similarity using siamese networks for audio-to-score alignment," in *Proc. IEEE Eur. Signal Process. Conf*, 2021, pp. 141–145.

[11] J. Devaney, "Estimating onset and offset asynchronies in polyphonic score-audio alignment," *J. New Music Res.*, vol. 43, no. 3, pp. 266–275, 2014.

[12] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Audio-to-score alignment using deep automatic music transcription," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2021, pp. 1–6.

[13] M. Miron, J. J. Carabias-Orti, and J. Janer, "Audio-to-score alignment at note level for orchestral recordings," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2014, pp. 125–130.

[14] X. Wang, R. Stables, B. Li, and Z. Duan, "Score-aligned polyphonic microtiming estimation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2018, pp. 361–365.

[15] R. Agrawal, D. Wolff, and S. Dixon, "Structure-aware audio-to-score alignment using progressively dilated convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2021, pp. 571–575.

[16] S. Wang, S. Ewert, and S. Dixon, "Robust and efficient joint alignment of multiple musical performances," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2132–2145, Nov. 2016.

[17] B. Li and Z. Duan, "An approach to score following for piano performances with the sustained effect," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2425–2438, Dec. 2016.

[18] N. Orio and D. Schwarz, "Alignment of monophonic and polyphonic music to a score," in *Proc. Int. Comput. Music Conf.*, 2001, pp. 155–158.

[19] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 185–188.

[20] F. Soulez, X. Rodet, and D. Schwarz, "Improving polyphonic and poly-instrumental music to score alignment," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2003, pp. 1–6.

[21] J. Devaney, M. I. Mandel, and D. P. W. Ellis, "Improving MIDI-audio alignment with acoustic features," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 45–48.

[22] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2385–2397, Nov. 2011.

[23] C. Joder, S. Essid, and G. Richard, "Optimizing the mapping from a symbolic to an audio representation for music-to-score alignment," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 121–124.

[24] C. Joder, S. Essid, and G. Richard, "Learning optimal features for polyphonic audio-to-score alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2118–2128, Oct. 2013.

[25] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno, "Incremental bayesian audio-to-score alignment with flexible harmonic structure models," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2011, pp. 525–530.

[26] C. Joder and B. W. Schuller, "Off-line refinement of audio-to-score alignment by observation template adaptation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2013, pp. 206–210.

[27] C. T. Chen, J. S. R. Jang, W. S. Liu, and C. Y. Weng, "An efficient method for polyphonic audio-to-score alignment using onset detection and constant Q transform," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2016, pp. 2802–2806.

[28] J. Syue et al., "Accurate audio-to-score alignment for expressive violin recordings," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2017, pp. 250–256.

[29] T. Kwon, D. Jeong, and J. Nam, "Audio-to-score alignment of piano music using RNN-based automatic music transcription," in *Proc. Sound Music. Comput. Conf.*, 2017, pp. 1–6.

[30] B. Niedermayer, "Improving accuracy of polyphonic music-to-score alignment," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2009, pp. 585–590.

[31] B. Niedermayer and G. Widmer, "A multi-pass algorithm for accurate audio-to-score alignment," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2010, pp. 417–422.

[32] T. M. Wang, P. Y. Tsai, and A. W. Y. Su, "Score-informed pitch-wise alignment using score-driven non-negative matrix factorization," in *Proc. IEEE Int. Conf. Audio, Lang. Image Process.*, 2012, pp. 206–211.

[33] S. Wang, S. Ewert, and S. Dixon, "Compensating for asynchronies between musical voices in score-performance alignment," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2015, pp. 589–593.

[34] Z. Liao, Y. Yu, B. Gong, and L. Cheng, "Audeosynth: Music-driven video montage," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 68:1–68:10, 2015.

[35] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[36] C. Liang, L. Su, Y. Yang, and H. Lin, "Musical offset detection of pitched instruments: The case of violin," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2015, pp. 281–287.

[37] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proc. Int. Soc. Music Inf. Retrieval.*, 2019, pp. 33–43.

Zhicheng Lian received the bachelor's degree in 2021 from Tianjin University, Tianjin, China, where he is currently working toward the master's degree with the College of Intelligence and Computing. His research interests include music information retrieval and audio processing.

Haonan Cheng (Member, IEEE) received the bachelor's and Ph.D. degrees in computer science from Tianjin University, Tianjin, China, in 2016 and 2021, respectively. She is currently an Assistant Professor with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China. Her research interests include cross-modal sound generation, audio processing, music information retrieval, computer graphics and virtual reality.

Jiawan Zhang (Senior Member, IEEE) received the master's and Ph.D. degrees in computer science from Tianjin University, Tianjin, China, in 2001 and 2004, respectively. He is currently a Professor with the College of Intelligence and Computing, Tianjin University. His research interests include visual computing and data visualization.