An Approach to Score Following for Piano Performances With the Sustained Effect

Bochen Li, *Student Member, IEEE*, and Zhiyao Duan, *Member, IEEE*

***Abstract*—One challenge in score following for piano music is the sustained effect, i.e., the waveform of a note lasts longer than what is notated in the score. This can be caused by expressive performing styles such as the legato articulation and the usage of the sustain and the sostenuto pedals and can also be caused by the reverberation in the recording environment. This effect creates nonnotated overlappings between sustained notes and latter notes in the audio. It decreases the audio-score alignment accuracy and robustness of score following systems and makes them be prone to delay errors, i.e., aligning audio to a score position that is earlier than the correct position. In this paper, we propose to modify the feature representation of the audio to attenuate the sustained effect. We show that this idea can be applied to both the chromagram and the spectral-peak representations, which are commonly used in score following systems. Experiments on the MAPS dataset show that the proposed method significantly improves the alignment accuracy and robustness of score following systems for piano performances, in both anechoic and highly reverberant environments.**

***Index Terms*—Audio-score alignment, reverberation, score following, sustain pedal.**

I. Introduction

THE two commonly used digital representations of music are the audio waveform and the musical score. While the audio provides rich information about the musical performance, it is not structured for computers to understand directly. The musical score (e.g., MIDI and music XML), on the other hand, is machine-readable, yet lacks the expressiveness of the musical performance. Aligning musical audio with the score has been an important research topic in music information retrieval since the 1990s [1].

An audio-score alignment algorithm can be classified as offline or online, according to its requirement on the audio input. An *offline* algorithm needs to access the whole sequence of the audio before starting the alignment process. An *online* algorithm (also called *Score Following* if runs in realtime) is able to align each audio frame with the score "without looking into the future" or "by only looking into the near future" (i.e., with some delays) of the audio performance. Since online algorithms require less information than offline algorithms, they are more challenging to design to achieve the same alignment accuracy and robustness. However, only online systems can support real-time applications if provided with enough computational resources.

Audio-score alignment has many applications. Offline algorithms have been used to construct multi-modal (e.g., video, audio, and score) music digital libraries [2], to build query-by-humming systems [3], to design piano tutoring and grading systems [4]. Online algorithms have been used in automatic music accompaniment for a live soloist [5], interactive piano pedagogy [6], real-time enhanced music enjoyment of orchestral performances [7], score-informed source separation [8], automatic coordination of audio-visual equipment [9], and automatic page turning [10]. Other inspiring potential applications include automatic display of lyrics or opera subtitles and the control of lighting and camera movement on stage.

In this paper, we focus on score following for piano performances. Piano is one of the most popular instruments in the world [11]. It is played worldwide in a variety of music genres including classical, jazz/blues, rock, and pop. It produces sounds ranging from monophonic (e.g., nursery rhymes) to highly polyphonic (e.g., piano arrangements of symphonies). It is also one of the few instruments that do not require accompaniment, thanks to its wide pitch range and highly polyphonic nature. However, the flexibility and expressiveness make piano performances difficult to follow by computers. More specifically, in this paper, we analyze the *sustained effect* in score following for piano performances (see Section III). This effect can be caused by the legato articulation of notes, the usage of the sustain and the sostenuto pedals, and also the reverberation of the recording environment. It extends the sound of notes longer than the expected length notated in the score, creating overlappings between sustained notes and latter notes, hence causing mismatch between audio and score.

We propose an approach to modify the feature representations of the audio to attenuate the sustained effect. We first perform onset detection and treat all frames within a region immediately after an onset as frames that potentially contain the sustained effect. We then analyze the spectra of the signal in these frames and detect spectral components that are considered as an extension from previous notes. We reduce the energy of these components in the audio representation to attenuate the sustained effect. We implement this basic idea for two commonly used audio representations in audio-score alignment, the chromagram and the spectral-peak representations. We test the modified representations within a hidden Markov model (HMM)-based score following framework [12]. Experiments on the MAPS dataset [13] show that the proposed approach greatly improves the alignment accuracy and robustness of highly expressive piano performances with different degrees of the sustained effect.

Manuscript received February 8, 2016; revised July 1, 2016 and August 16, 2016; accepted September 4, 2016. Date of publication September 20, 2016; date of current version October 23, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hiroshi Saruwatari.

The authors are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (e-mail: bochenli@rochester.edu; zhiyao.duan@rochester.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2611938

It is noted that our proposed sustained-sound reduction operation does not always reduce the sustained effect correctly. First, some frames may not have the sustained effect, yet we perform the operation in all frames within a region after an onset. Second, we cannot discriminate whether the extension of a note is due to the sustained effect or is because the note is notated that way. Third, onsets may be wrongly detected. In all these three cases, the sustained-sound reduction operation will cause new audio-score mismatch. However, in Section IV-C we will see that only in the last case the operation is harmful to the alignment, while in the other two cases the operation is still helpful or has no significant effect.

A preliminary version of this work has been published in [14] to deal with the sustained effect caused by the sustain pedal usage in piano score following. A method to reduce the effect using the spectral peak representation of audio was proposed. In this paper, we extend the concept of the sustained effect to legato note articulation and reverberation. We further develop the sustained-sound reduction operation for the chromagram representation of audio. We also add detailed discussions of the effects of the operation, especially when they are wrongly applied. We conduct systematic experiments on piano pieces with different degrees of the sustained effect, compare the proposed method with baselines, and analyze the influences of key parameters.

The rest of the paper is organized as follows. We first introduce related work in Section II. We then illustrate several specific properties of piano music in Section III to set up the background for the proposed approach. In Section IV, we propose a sustained-sound reduction operation with an online onset detection technique to reduce the sustained effect in audio representations, and also discuss the influence when the operation is wrongly applied in three cases. Systematic experiments are illustrated in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

A. Early Work

Audio-score alignment has been an active research topic for two decades. Early approaches focus on monophonic audio, such as vocals and monophonic instrumental solos [15]–[17]. For polyphonic music, Orio and Schwarz [18] firstly applied Dynamic Time Warping (DTW) for alignment on string and wind ensembles. This method is computationally demanding because it considers alignment between every pair of audio and score segments. Müller *et al.* [19] constrained the alignment within a region found through a multi-scale analysis. They obtained similar alignment results but with 50 times less memory cost. Kapryknowsky and Rodet [20] proposed a short-time DTW approach to reduce the computational complexity of standard DTW.

B. Online Alignment

An online audio-score alignment system typically contains two modules: a processor and an observer. The *processor* (or *process model*) is a hypothesis generator. It continuously generates alignment hypotheses for the incoming audio frame. The *observer* (or *observation model*) is a hypothesis evaluator. It evaluates the alignment hypotheses by matching the audio frame with the hypothesized score positions and chooses the best hypothesis. The contribution of this paper lies in the observer module.

One type of commonly used processors is the online DTW algorithm [21]. Alignment hypotheses generated by this method are neighboring cells within an adaptive band centered around the current cell in the alignment matrix. This method, however, has no guarantee of the continuity of the alignment path without retrospective adjustments. This idea is further developed by the "backward-forward strategy" to reconsider the past decisions [10], and the incorporation of a tempo model [22] for robustness. Another processor proposed in [16] employs stochastic models, where the score position hypotheses are represented by a probability density function. The benefit of this method is that the random factors in the system can better cope with the uncertainties in real performances. Similarly, Duan and Pardo [8] proposed a hidden Markov process model in a 2-D continuous state space with the score position and tempo being the two dimensions. Pardo and Birmingham [23] modeled score forms to deal with large-scale structural variations such as skipping or repeating of a section.

The key problem of designing an observer is the choice of representations of audio and score. The most commonly used representation is chromagram [7], [24]–[28], which well describes the harmonic progression of music. It can be calculated for both audio and score, and their match can be evaluated using Euclidean or cosine distances. Müller and Ewert [29] discussed ways of enhancing and implementing chroma features. A similar but richer representation is semigram. It also uses a semitone frequency scale but does not fold different octaves into one as the chromagram does. Semigram was first employed by Dixon [21] in score following, and was later adopted by other systems [30], [10] and [22]. Another commonly used audio representation is the spectral-peak representation [8], [12], [18]. Spectral peaks are ideally caused by harmonics of notes, hence they convey pitch information [31], through which the match between audio and score can be defined. Other representations include auditory filter bank responses [32], Nonnegative Matrix Factorization (NMF)-based multi-pitch representation [33], [34], and adaptive-template-based observation models [35]. Onsets are also modeled in some representations [26], [36] to achieve more accurate alignment between audio and score.

C. Score Following for Piano Music

Few methods have been proposed to address score following specifically for piano music. General score following systems may achieve better results on piano pieces than other instruments in the offline scenario, thanks to the clear onsets and consistent timbre of piano notes [30]. In online scenarios, however, this advantage is likely to be dominated by the disadvantages of the high loudness contrast of simultaneous notes and the sustained effect caused by the legato articulation of notes and the usage of the sustain and sostenuto pedals.

The sustained effect for audio-score alignment was firstly observed by Orio and Schwarz [18]. They stated that partials of the previous notes can be still present in the beginning of the next notes, due to the legato articulation or reverberation. To cope with this, they used the first-order positive difference of magnitude spectrum (spectral flux) as the audio representation instead of the original spectrum to emphasize onsets. This was also adopted in [10], [21], [30]. The problem of this representation is that in most inter-onset frames the spectral flux is close to zero, because their spectra are very similar to those of their previous frames. This undermines the inference of the score position in these inter-onset frames, and results in outliers in the alignment path found by the online DTW algorithms, as described in [30].

It is worth to mention Niedermayer *et al.*'s finding on the influence of the sustain-pedal usage on audio-score alignment accuracy [37]. They compared alignment results on the same music pieces performed with and without the sustain pedal, and found that the influence was negligible. However, they explained that this might be due to the rare usage of the sustain pedal in the pieces that they tested, which were all Mozart pieces. In fact, the sustain pedal was rarely used before the Romantic era but has been commonly used since then (see Section III-B). Another reason for Niedermayer *et al.*'s observation, we argue, is that the algorithm used for evaluation was an offline algorithm, which is more robust to the local mismatch between audio and score as a global alignment is employed. For online algorithms, however, they are more sensitive to local audio-score mismatch, which can be greatly introduced by the the usage of the sustain pedal.

III. Sustained Effect in Piano Music

In this section, we start from basic acoustical properties of piano notes, then we describe in detail the sustained effect, its major causes, and its influences on score following.

A. Acoustical Properties of Piano Notes

Each piano key is associated with a hammer, one to three strings, and a damper that touches the string(s) by default. When a key is depressed, its hammer strikes the string(s) while the damper is released from the string(s). This yields an impulse-like articulation at the note onset. The loudness of the note is determined by the velocity of the hammer strike, which is affected by how hard the key is depressed. The string(s) then vibrate freely until the damper returns to the string(s) when the key is released. The free vibration of the string(s) produces an exponential energy decay of the waveform. The pitch of the note, however, does not change. It is determined by the material, length, and tension of the string(s) and is pre-tuned. A performer cannot control the pitch (e.g., playing vibrato) as one could do on other instruments (e.g., strings, winds). To summarize, there are three important acoustical properties of piano notes as follows. In Section IV, we will show how these properties can be leveraged to design an approach to reducing the sustained effect.

1) *Strong Onset*: Piano note onsets are generally easier to detect thanks to the impulsive articulation.

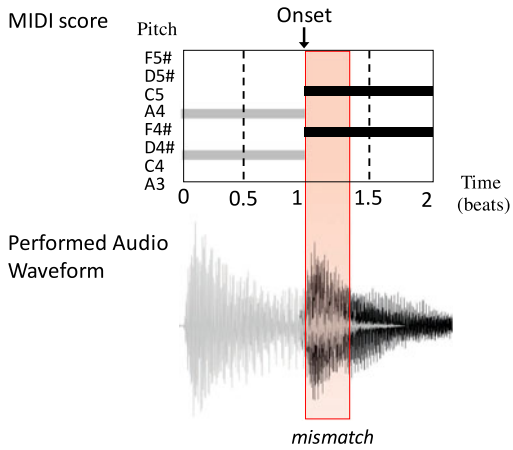


Fig. 1. Illustration of the sustained effect and the audio-score mismatch problem it causes. The gray notes are extended in the audio waveform longer than their notated length in the score.

2) *Energy Decay*: The decay time varies for different strings, and can be up to 10 seconds if the strings vibrate freely.
3) *Constant Pitch*: The pitch of a note is constant and cannot be controlled during the entire process of a note.

B. Sustained Effect and Its Major Causes

The *sustained effect* refers to the phenomenon that a note is sustained longer than its notated length in the score. It is very common in piano performances. Fig. 1 shows the concept. According to the score, the two gray notes (A4 and D4) should stop at Beat 1. Their waveforms in the audio performance, however, are sustained longer and are blended with the latter two black notes. This causes mismatch between the audio and the score in the shaded region at the beginning of the two black notes. The length of this region depends on when the gray notes stop sounding. Statistical analysis of this region length in piano performances is presented in Section V-A1. In the following, we analyze the several major causes of it.

1) Legato Articulation: The musical term *legato* indicates that notes are played smoothly and connected. For piano, the general practice for producing legato notes is to release a key after depressing the key for the following tone with an overlapping [38]. To measure the degree of legato articulation, Repp [39] introduced the Key Overlap Time (KOT) for adjacent tones, which is defined as "the time interval between the onset of key depression for one tone and the key release for the preceding one" [39]. Although the note transition is still not very smooth given the strong onsets of piano notes, this practice does attenuate the percussive sensing.

2) Sustain/Sostenuto Pedal Usage: Modern pianos generally have three foot pedals: sustain, sostenuto, and soft pedals; some models omit the sostenuto pedal. When the sustain pedal is pressed, all dampers of all notes are raised from all strings, no matter whether a key is depressed or released. Therefore, its usage will sustain all notes whose keys are released when the pedal is being pressed. The sostenuto pedal behaves similarly, but only keeps raising dampers that have already been raised

without affecting others. Therefore, its usage will sustain notes that are activated before the pedal is pressed *and* that are released while the pedal is being pressed. The soft pedal changes the way that the hammer strikes the string(s), hence it affects the timbre and loudness, but its use is rare compared to the use of the other pedals.

The sustain and sostenuto pedals, especially the sustain pedal, have been commonly used since the Romantic era in Western music history, and in modern piano performances of many different styles. Fig. 2(a) shows the proportions of pieces with different degrees of sustain-pedal usage in the MAPS dataset [13], which is a commonly used piano dataset for music transcription and score following with a good coverage of composers in different eras. The usage of the sustain pedal produces round and velvety timbre for lyrical expressions. It also lets pianists sustain notes which would otherwise be out of reach. It can also accomplish legato passages which would otherwise have no possible fingering. While some composers and music arrangers use pedal marks to notate it, appropriate use of the sustain pedal is more often left to the performers' discretion. In addition, the pedals can be partially pressed, causing a slighter sustain effect as the dampers slightly touch the strings. A detailed analysis of the sustain-pedal effects can be found in [40].

3) Reverberation: While legato articulation and pedal usage depend on the piece that is being performed, room reverberation is a universal cause of the sustained effect in any real-world piano performances. In this case, all notes are sustained. Because reverberation can be well modeled as a Linear Time-Invariant (LTI) system, it does not change the pitch of the notes. The extent of the sustained effect is determined by the architecture design (e.g., size, shape, material). It can be measured by reverberation time RT_{60}: the time it takes for the room impulse response to decay 60 dB from the original level. It ranges from about 0.5 seconds (e.g., practice room or studio) to 3 seconds (e.g., church or large concert hall) [42]. Fig. 2(b) shows the reverberation time of three well-known concert halls.

C. Influences of Sustained Effect on Score Following

The sustained effect causes non-notated overlappings between sustained notes and new coming notes, which influence score following for piano performances. Take Fig. 1 as an example, when the audio has entered into the shaded region, the score follower may still match to a position before Beat 1. This is because the audio, which contains both the gray and black notes, does not show a much better match with the shaded region of the score than with a location before Beat 1. Therefore, the follower often wrongly aligns to the positions before the correct one (i.e., *delay errors*). This may cause considerable lags in the alignment, and affect the score following accuracy and robustness when such errors are accumulated.

IV. Proposed Approach

In this section, we propose an approach to address the sustained effect in score following for piano performances. We first locate the regions that contain potential mismatch between audio and score caused by the sustained effect through online onset detection. We then propose an operation in two kinds to reduce the sustained sound in the chromagram and spectral-peak representations of the audio, respectively. Note that not all note onsets are followed by a region containing the sustained effect, so the operation can be wrongly applied. We analyze these cases and the consequences of the wrongly applied operations. Finally, we integrate these operations into an online score following system.

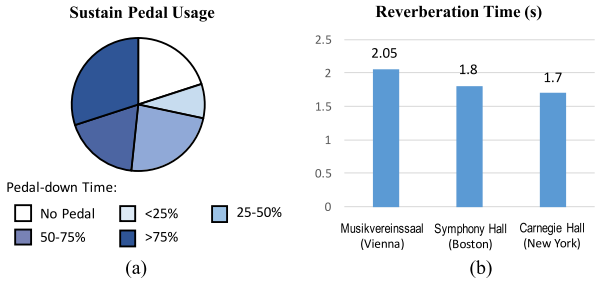


Fig. 2. Statistics of two causes of the sustained effect. (a) Distribution of the 60 acoustic pieces in the MAPS dataset [13] according to the degree of pedal usage. Pedal-down time denotes the percentage of the performing time when the sustain pedal is depressed. (b) Reverberation time of three famous concert halls: Musikvereinsaal in Vienna, Symphony Hall in Boston, and Carnegie Hall in New York, obtained from [41].

A. Online Onset Detection

The sustained effect, when it appears, always appears in an audio region right after the onset of new notes, e.g., the shaded region in Fig. 1. However, not all onsets are followed by the sustained effect. For example, there might be no sustained effect at all between staccato notes. Nevertheless, we propose to locate all the potential audio regions through online onset detection and apply the sustained-sound reduction operation. The operation may be wrongly applied, but we will see that its consequences are minimal if the onsets are correct in Section IV-C.

Onset detection for music signals is a well-studied topic [43], [44]. A general framework is to first convert audio features (e.g., spectral flux) into a detection function through several signal processing steps or machine learning modules (e.g., neural networks [45]), and then detect onsets through normalization, thresholding, and peak picking. Most methods work in an offline fashion, but there exist online adaptations of these methods [46]. In this paper, we propose an online adaptation of a spectral-flux-based onset detection method employing several simple signal processing steps, in favor of simplicity and efficiency.

We assume that the volume of the input piano performance has been normalized. In practice, this normalization can be adjusted during the system setup. Here we simply normalize its Root Mean Square (RMS) value to 1 before the online processing starts. Then let $\mathbf{Y}(n,k)$ be the Short-Time Fourier Transform (STFT) magnitude spectrogram of the music signal calculated with a 46.4 ms Hamming window and a 10 ms hop size, where n and k are time frame and frequency bin indices, respectively. The first step is a logarithmic compression with a ratio $\gamma = 0.2$ of the spectrogram to enhance the high-frequency content:

$$\tilde{\mathbf{Y}}(n,k) = \log\left(1 + \gamma \cdot \mathbf{Y}(n,k)\right). \tag{1}$$

This compression step yields better results because high frequency components are more indicative of note onsets but are relatively weak in linear-amplitude spectrograms [47].

The second step is to calculate the spectral flux as an onset salience function $S(n)$, which is the positive first-order difference between consecutive frames of the enhanced spectrogram across all frequency bins:

$$S(n) = \sum_k \left| \tilde{\mathbf{Y}}(n,k) - \tilde{\mathbf{Y}}(n-1,k) \right|_{\geq 0}, \tag{2}$$

where $|\cdot|_{\geq 0}$ denotes half-wave rectification, i.e., keeping non-negative values while setting negative values to 0.

The third step is to enhance the salience function to account for loudness variations. This is often done by subtracting the salience function by a smoothed version of itself [44]. For our online setting, the smoothed version is calculated from past frames [46]:

$$\tilde{S}(n) = \frac{1}{\omega+1} \cdot \sum_{m=-\omega}^{0} S(n+m), \tag{3}$$

where ω is the sliding window size for the local average calculation. Then the *enhanced onset salience function* $\hat{S}(n)$ can be calculated by taking the positive difference between the original salience function and its smoothed version:

$$\hat{S}(n) = |S(n) - \tilde{S}(n)|_{\geq 0}. \tag{4}$$

The last step is to determine onsets from the enhanced salience function. For offline methods, this is often achieved by peak-picking [43]. For online methods, however, peak-picking requires at least one frame of delay. In our approach, in order to avoid inherent delays, we opt to employ another approach inspired by [46]. We set an amplitude threshold α and a time threshold β. We report an onset in the current audio frame if three conditions are all satisfied: 1) the enhanced salience $\hat{S}(n)$ is greater than α; 2) the enhanced salience is greater than the salience in all past ω frames, where ω is the window size in Eq. (3); 3) no other onsets have been reported in the past β frames. This approach ensures real-time onset detection without any delays, but often reports onsets earlier than the true onsets. In our experiments, about 30% of correctly detected onsets (i.e., those deviating less than three frames from ground-truth) are in the same frame as ground-truth. More than 60% are earlier and less than 10% are later. Compared with a delayed detection using peak-picking, this online approach is still preferred in our score following system.

Among the parameters α, β, and ω, the amplitude threshold parameter α has the largest impact on onset detection, specifically on the ratio between false negatives (miss errors) and false positives. A miss error may lead to the miss detection of a sustained-effect region, while a false positive will break an inter-onset interval into two and wrongly apply the sustained-sound reduction operation in the second half. These errors will have different effects on the score following results. In Section IV-C3 we discuss how false positives affect score following. In Section V we experimentally investigate the impact of α on onset detection and the final alignment performance.

B. Sustained-Sound Reduction

In this subsection, we propose an approach to reduce the sustained sound in the audio representation of all frames in the audio regions detected in the previous subsection. While the length of the shaded region depends on the degree of the sustained effect, in the experiments we simply use a unified size of $L = 150$ ms or the interval to the next onset if it is within 150 ms. We conduct experiments to show the sensitivity of the parameter L in Section V-C2.

The basic idea is to inspect the spectrum and reduce the spectral components that are sustained from previous notes. There are two main tasks: 1) Identify components that are sustained, and 2) reduce these components in the audio representation. Answers to these questions depend on the audio representation that is used to match the audio with the score. In the following, we propose methods for two commonly used audio representations in score following: the chromagram and the spectral-peak representations.

1) Spectral Subtraction for Chromagram Representation: The chromagram [48] can well represent the harmonic content of the music audio and is less sensitive to timbral variations than the spectrogram. It has been commonly used in audio-score alignment approaches [7], [24]–[28]. One way to calculate a 12-d chroma vector $\mathbf{Ch}_a$ for an audio frame is from the STFT magnitude spectrum using weighting functions [49]. Each dimension of the chroma vector is a weighted sum of the energy of all frequency bins in the spectrum. Each weighting function is a mixture of Gaussians that are centered at frequencies of the pitch class at different octaves using the standard A = 440 Hz tuning and a standard deviation of a quarter-tone.

To deal with the sustained effect in a detected region after an onset, we calculate the chroma vectors from a modified spectrogram $\mathbf{Y}^*(n,k)$ instead of the original magnitude spectrogram $\mathbf{Y}(n,k)$, by subtracting a reference spectrum $\mathbf{Y}(m,k)$:

$$\mathbf{Y}^*(n,k) = |\mathbf{Y}(n,k) - f(\hat{S}(n)) \cdot \mathbf{Y}(m,k)|_{\geq 0}. \tag{5}$$

The reference frame m is set to 5 frames (i.e., 50 ms) before the onset frame in our implementation. The half-wave rectification prevents $\mathbf{Y}^*(n,k)$ from being negative after subtraction. $f(\cdot)$ is a confidence function of onset detection controlled by the enhanced onset salience $\hat{S}(n)$, and is defined as:

$$f(x) = \begin{cases} 0 & \text{if } x < \alpha \\ \frac{1}{\alpha'-\alpha}(x-\alpha) & \text{if } \alpha \leq x \leq \alpha' \\ 1 & \text{if } x > \alpha' \end{cases}. \tag{6}$$

where α is the threshold for onset detection in Section IV-A, and α' is another conservative threshold which decides when the full amount of the reference spectrum is subtracted. When the salience $\hat{S}(n) > \alpha'$, the onset detection is confident enough and the full amount of the reference spectrum is subtracted to enhance new notes. When $\alpha \leq \hat{S}(n) \leq \alpha'$, the detected onset might be a false positive. We subtract just a portion of the full amount so that the chromagram does not become totally blank in that case. Note that ideally the reference frame m would serve our purpose the best if it were immediately before the onset, even though the spectral difference among the several frames before

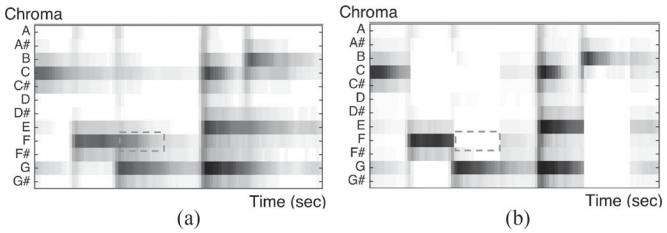


Fig. 3. Chromagrams calculated without (a) and with (b) spectral subtraction. Sustained sounds after onsets (e.g., that marked by the rectangle) are reduced by the sustained-sound reduction operation.

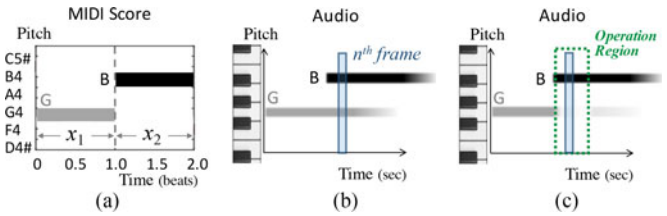


Fig. 4. Illustration of the audio-score mismatch reduced by the proposed sustained-sound reduction operation, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The blue patch in (b) and (c) indicates the current frame (n-th frame) which lies in the operation region.

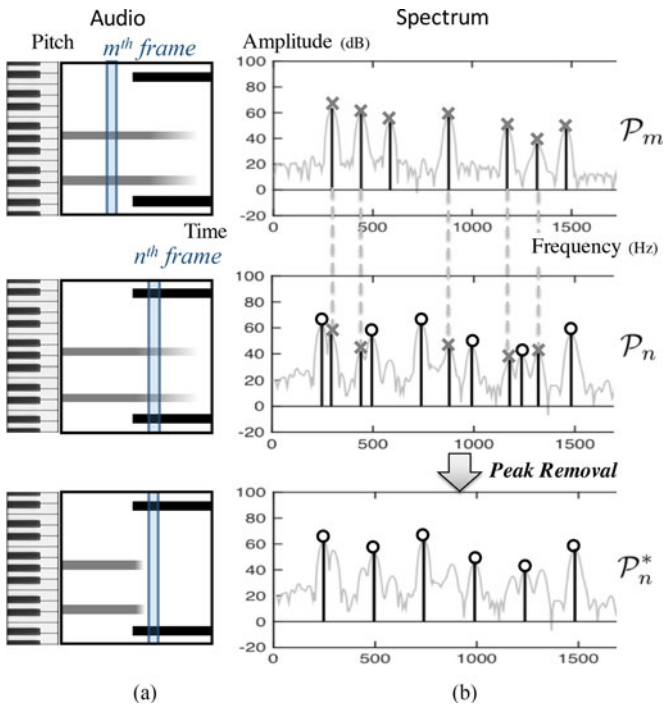


Fig. 5. Illustration of the spectral peak removal idea. The m-th frame is the reference frame and the n-th frame is a frame under the removal operation. (a) Audio representations in the pianoroll format before and after peak removal. (b) Magnitude spectra with detected spectral peaks in the m-th and n-th frames. Peaks marked by crosses correspond to the first two notes. Peaks marked by circles correspond to the latter two notes.

the onset is subtle. We pick m as five frames before to leave a safe margin in case the detected onset is a delayed prediction, while it is still not too large to reach the previous onset.

Fig. 3 compares the chromagrams calculated with and without the spectral subtraction operation in Eq. (5). With the spectral subtraction operation, spectral components of sustained sound are greatly reduced (instead of totally removed) from frames right after onsets, whereas components of the new notes remained. This greatly reduces the audio-score mismatch as illustrated in Fig. 4. It shows a MIDI score with two inter-onset segments x_1 and x_2, where note G is supposed to end when note B starts. For an audio frame right after the onset (e.g., the n-th frame), we can see that it contains both notes, including the extension of note G due to the sustained effect. It is therefore not a precise performance of the correct segment x_2 in the score. Let $M(y_n, x_1)$ denote the match between the audio frame y_n and the score segment x_1, and let $M(y_n, x_2)$ be defined similarly, then it is difficult to tell which match is better. After the sustained sound reduction operation, however, the note G is greatly reduced in the audio representation of the n-th frame, and it becomes clear that it has a better match to the correct score segment x_2.

2) Peak Removal for Spectral-Peak Representation: The spectral-peak representation is also commonly used in audio-score alignment [8], [12], [18]. In this representation, the magnitude spectrum is reduced to a set of frequency-amplitude pairs of significant peaks:

$$\mathcal{P} = \{\langle f_i, a_i \rangle\}_{\mathrm{i}=1}^{K}, \tag{7}$$

where K is the total number of peaks detected in the frame. The basis of this representation is that ideally the peaks correspond to harmonics of notes, through which the match between audio and score can be evaluated. For example, a good match would be that the score contains notes whose harmonics appear and only appear at the peaks.

To reduce the sustained effect in a detected region, we propose to remove peaks that correspond to sustained sounds in the spectral-peak representation. Fig. 5 illustrates the idea. For each frame in a detected region (e.g., the n-th frame), we compare its spectral peaks with those in a reference frame before the onset (e.g., the m-th frame), and remove peaks that seem to be extended from the earlier frame. Let $\mathcal{P}_m = \{\langle f_i^m, a_i^m \rangle\}_{\mathrm{i}=1}^{K_m}$ be the total K_m peaks detected in the m-th frame, and $\mathcal{P}_n = \{\langle f_j^n, a_j^n \rangle\}_{\mathrm{j}=1}^{K_n}$ be the total K_n peaks detected in the n-th frame. A peak in the n-th frame whose frequency is very close to *and* whose amplitude is smaller than those of a peak in the m-th frame is considered as an extension and is removed. Note that repeated notes will not be removed in this way as the amplitude criterion is not met. Thanks to the same reason, partials of the new notes are generally not removed even if they overlap with partials of the extended notes, because a significant energy increase is often observed at the onset of the new notes. After peak removal, a new spectral peak representation of the n-th frame is obtained as

$$\mathcal{P}_n^* = \mathcal{P}_n - \left\{\langle f_i^n, a_i^n \rangle : \exists j \text{ s.t. } |f_i^n - f_j^m| < d, a_i^n < a_j^m \right\}, \tag{8}$$

where d is the threshold for the allowable frequency deviation, which is set to a quarter-tone in this paper. Although this

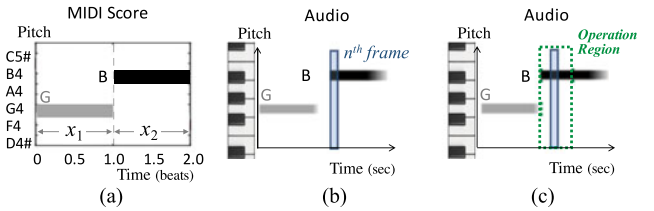


Fig. 6. Illustration of the negligible effect of the proposed sustained-sound reduction operation when there is no sustained sound, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.

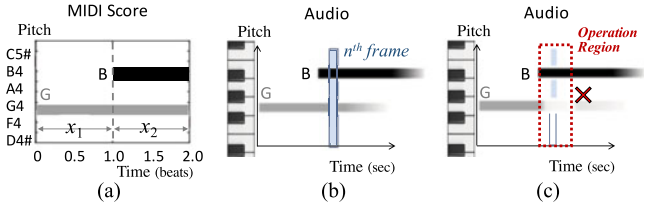


Fig. 7. Illustration of the new audio-score mismatch introduced by the proposed sustained-sound reduction operation when the sustained sound is not due to the sustained effect but is notated in the score, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.

operation is not applied to all frequencies as the spectral subtraction operation does for the chromagram representation, it also reduces the audio-score mismatch and reduces the delay error in score following.

C. What If the Proposed Operations Are Wrongly Applied?

As described in Section IV-B, the proposed sustained-sound reduction operation is applied to the entire 150-ms region after an onset. It reduces audio-score mismatch due to the sustained effect in that region and reduces delay errors in score following. However, not every region after an onset contains the sustained effect. In addition, onset detection has false positive errors. In these cases, the proposed operation will be wrongly applied. Will it be harmful for the audio-score match and score following? There are in total three cases in which the operation will be wrongly applied, and we analyze them in detail in this section.

1) Case 1: Onset Is Correct but There Is No Sustained Sound: In this case, the onset is correctly detected but there is no sustained sound after the onset, i.e., the old notes (e.g., staccato notes) simply cease before the next onset. Fig. 6 shows an example. The spectral subtraction operation for the chromagram representation in Eq. (5) will still subtract a portion of the old notes' spectrum and may contaminate the spectrum of the new notes. However, this effect is subtle. The subtraction is half-wave rectified, so it will not affect frequencies that the new notes do not contain. It will not affect the frequencies that the old notes do not contain either, since there is nothing to subtract. The only frequencies that it affects are frequencies shared by the old and new notes. In these frequencies, the new notes are likely to have a larger amplitude than the old notes as they have just started. Therefore, the subtraction will just reduce the amplitude instead of removing them. This corresponds to a slight timbre change in the modified audio representation, but the same harmonic content is represented, which is the key for audio-score match in the chromagram representation.

For the peak removal operation, no partials of the new notes, whether overlapped with the old notes or not, are likely to be affected. This is because the amplitude criterion in Eq. (5) is not likely to be satisfied due to the amplitude increase at the onset, as explained in Section IV-B2. Therefore, the proposed operation in both kinds are not harmful to the audio-score match nor score following. However, if the operation region is too long, the amplitude criterion might be satisfied in some latter frames and there can be some slight negative effects. The above

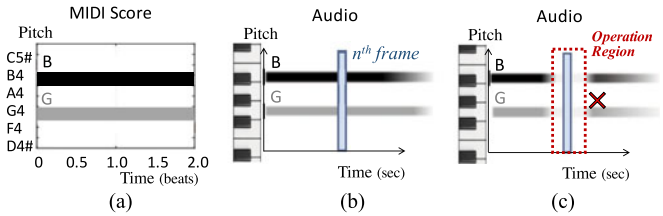


Fig. 8. Illustration of the proposed sustained-sound reduction operation when the onset is a false positive, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The audio representation in the rectangular region is likely to preserve some energy of the notes after the spectral-subtraction operation but can be totally blank after the peak-removal operation.

analyses are verified by the comparable results achieved with and without the proposed operation on pieces that do not contain apparent sustained effect and the experiment on the operation region length in Section V.

2) Case 2: Onset Is Correct but Sustained Sound Is Notated: In this case, the onset is correctly detected, and there is sustained sound after the onset. However, this sustained sound is not due to the sustained effect but simply because the notes are supposed to sustain according to the score. This case happens very often in piano performances, e.g., the right hand is playing fast melody while the left hand is holding long chords. The left-hand notes are sustained according to the score. Fig. 7 shows an example, where note G should be extended after the onset of note B according to the score. In this case, our proposed operation will wrongly remove note G in the n-th frame and decrease the match between the frame with the correct score segment x_2. However, the match between the n-th frame and the wrong score segment x_1 will be decreased even more. In fact, they will not match at all as they will not share any note after the operation. Therefore, the operation will actually make the score follower favor the correct segment x_2, even though it introduces new mismatch between audio and score. Good experimental results of the proposed approach on a variety of piano pieces support our claim.

3) Case 3: Onset Is A False Positive Error: The last case is that the onset is a false positive, and the sustained sound reduction operation is applied in frames that are in the middle of notes. Fig. 8 shows an example. In this case, for the spectral subtraction operation designed for the chromagram representation, the n-th frame is not likely to be severely affected. This is because the spectrum of the n-th frame is likely to be just

reduced but not totally removed, thanks to the confidence function employed in subtraction in Eq. (5). For the spectral-peak representation, however, almost all peaks in the n-th frame will be removed, because both the frequency and amplitude criteria in Eq. (8) for peak removal are likely to be satisfied for these decaying partials. Therefore, the spectral-peak representation can be severely contaminated and mismatch between audio and score can be introduced. The score follower can have difficulty in matching the audio to any score position. To sum up, false positive errors (which lead to lower precision) of onset detection are not significantly harmful for score following with the proposed sustained-sound reduction operation with the chromagram representation, but it is greatly harmful with the spectral-peak representation. Experiments in Section V-C1 support our claims.

Note that here we only analyze false positive errors but not miss errors in onset detection. This is because miss errors only prevent the proposed sustained-sound reduction operation from being applied but do not wrongly apply it. In addition, we do not consider offset detection for two reasons: 1) offsets are difficult to detect due to the lack of abrupt changes in energy and spectral content in the audio signal; 2) offsets are related to possible sustained effect in the past, hence their detection is not helpful in an online system.

D. The Score Following Framework

The modified chromagram and spectral-peak representations in Section IV-B can be employed by various score following frameworks to cope with the sustained effect in piano performances. In this work, we adopt an effective Markov model-based framework [12]. This framework uses a 2-d state variable s_n to represent the score position and tempo of the audio in the n-th frame. A process model $p(s_n|s_{n-1})$ is defined to describe how the states transition from one to another: The score position advances from the previous frame according to the tempo and the tempo changes through a random walk. An observation model $p(y_n|s_n)$ is defined to represent the likelihood of the hidden state s_n in explaining the observed spectrum y_n. Provided the process model and the observation model, the hidden state s_n can be inferred by particle filtering from current and previous audio observations $y_1, \ldots, y_n$. The framework is illustrated in Fig. 9.

The proposed audio representations are integrated into the observation model $p(y_n|s_n)$. For the chromagram representation, $p(y_n|s_n)$ is defined as a Gaussian distribution of the cosine distance between chroma vectors calculated from the audio and the score. For the spectral-peak representation, $p(y_n|s_n)$ is defined through a multi-pitch likelihood function proposed in [31]. Details of these observation models can be found in [12].

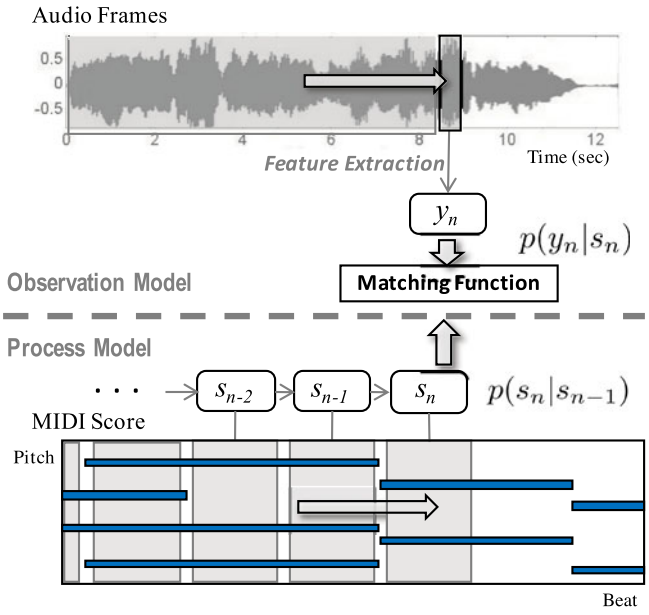


Fig. 9. The score following framework adopted to implement the proposed sustained-sound reduction approach.

V. Experiments

A. Experimental Set-Up

1) Dataset: To evaluate audio-score alignment systems, a set of audio recordings and their musical scores, as well as the ground-truth alignment between the audio and score are needed. For the audio recordings, we used the entire 60 acoustically recorded pieces from the two folders (*close* and *ambient*) of the MAPS dataset [13] to cover a large variety of styles, playing techniques, and the degree of the sustained effect in piano performances. These pieces were from more than 14 different composers including Chopin, Mozart, Liszt, etc., and contained different degrees of the sustain effect. All the pieces were acoustic recordings of a Yamaha Disklavier that took pre-generated MIDI performances as input. Since the degree of the sustained effect varies much even within the same piece, for each piece we choose a short clip with a length between 30 and 60 seconds in our experiments, aiming for a more similar degree of the sustained effect within each clip.

For the musical scores, we downloaded the standard MIDI score for each piece from the Classical Piano MIDI Page[1]. Beside the tempo differences, these MIDI scores also contain other minor differences from the MIDI performances in the MAPS dataset. These differences were due to the occasional missed or added notes, slight desynchronizations between melody and accompaniment [50], and different renderings of trills in the MIDI performances. Since the MIDI performance of each piece has exactly the same timings as the audio recording, we performed an offline DTW followed by manual corrections to align the MIDI performance with the MIDI score to obtain the ground-truth audio-score alignment.

According to the degree of the sustained effect, we categorized the 60 pieces into three groups. As the sustain pedal usage is a main source of the sustained effect, we first categorized the 12 pieces that were performed without any sustain pedal usage into the first group, *P1* (No Pedal). The pedal usage information was extracted from the MIDI performances using java

[1]http://piano-midi.de/. Downloaded in October, 2015.

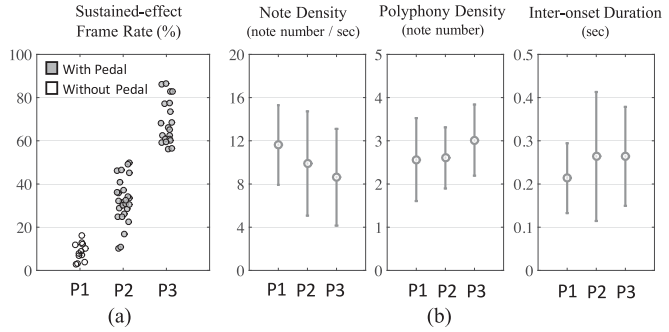


Fig. 10. Statistical MIDI information of the testing dataset. (a) Scatter plot of Sustained-effect Frame Rate of the three testing groups, calculated from MIDI performances. Each dot represents one piece. (b) Three measurements of piece complexity of the three testing groups, calculated from MIDI scores. The central circles and vertical bars denote the means and standard deviations.

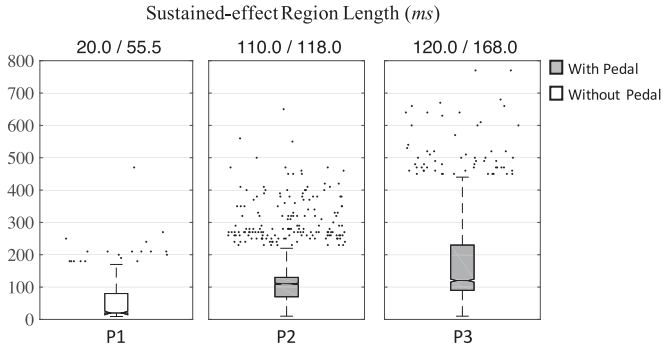


Fig. 11. Boxplot of the Sustained-effect Region Length (SRL) for all unique note onsets (excluding those with a zero SRL.) in pieces in *P1* (283 onsets), *P2* (1680 onsets), and *P3* (2275 onsets). Outliers are displayed in a more dispersed way for better visualization. Numbers above each box show the median/mean value of all the points.

MIDI tools [51]. Note that slight sustained effect may still exist in this group, due to the legato note articulations. For the rest 48 pieces, we then calculated the *Sustained-effect Frame Rate (SFR)*, which is defined as the percentage of mismatch frames due to the sustained effect, i.e., the *active notes* in the MIDI performance are more than those in the aligned MIDI score. Here the offsets of notes in the MIDI performance are extended from a sustain-pedal-down period to the next pedal-release time. The 48 pieces were then divided into two groups, *P2* (Slight) and *P3* (Heavy), by a threshold (50%) on the SFR. This threshold was chosen as it is at the low-density region of the distribution and it roughly balances the two groups. Fig. 10(a) shows the SFR of all the three groups. One may notice the overlap between *P1* and *P2* and ask why we did not use another threshold on SFR to divide *P1* and *P2*. Our rationale was that SFR is a rather arbitrary measure of the sustained effect while using pedal or not is a clear distinction among pieces. We would like to hold a set of pieces with absolutely no pedal usage.

Fig. 10(b) also shows statistics of three other measurements of the complexity of the pieces calculated from the MIDI score barring the sustain-pedal usage:

1) *Note density*: Average number of notes per second, N_{note}/T, where N_{note} is the total number of notes and T is the length of the piece.
2) *Polyphony density*: Average number of simultaneous notes, $\frac{1}{N}\sum_n u_n$, where N is the total number of frames and u_n is the number of active notes in frame n.
3) *Inter-onset duration*: Average time gap between the note onsets, $\frac{1}{N_{\text{onset}}}\sum_i (t_i - t_{\text{i}-1})$, where N_{onset} is the total number of unique onsets, and t_i is the time instant of the i-th unique onset.

We performed two-sample t-tests for all group pairs, and found that for all the three measurements, the groups are not significantly different at the significance level of 5%, except for the note density between *P1* and *P3*, which shows a p value of 0.0485. This helps us to rule out factors other than the degree of the sustained effect that may affect the score following performance, in the situation where controlled experiments are not easy to design.

To measure the length of each region with the sustained effect (the shaded region in Fig. 1) in our test pieces, we calculated the *Sustained-effect Region Length (SRL)* for the three groups in Fig. 11. SRL is defined as the length of the mismatch region between the MIDI performance and the aligned MIDI score for each unique onset. Again, the offsets of notes in the MIDI performance are extended from a sustain-pedal-down period to the next pedal-release time. Note that in Fig. 11 we exclude notes with a zero SRL, e.g., staccato articulations.

We further created another set by adding reverberation to all the 60 pieces by convolving each piece with a room impulse response ($RT_{60} = 2.34$ s) sampled from the St. Albans Cathedral[2]. Note that we did not use the reverberant pieces provided in the MAPS dataset for two reasons: 1) They were synthesized from MIDI using software instead of being acoustically recorded; 2) They use different pieces from the above-mentioned 60 acoustic recordings hence are hard to compare. By adding the external reverberation, we formed a controlled experiment.

2) Evaluation Measures: We use two kinds of evaluation measures for score following:

1) *Align Rate (AR)*: It is defined as the percentage of correctly aligned unique onsets among all unique onsets in the score [52]. Simultaneous onsets of different notes in the score are treated as a single onset. An onset i is considered correctly aligned if its aligned audio time $\hat{t}_i$ deviates less than a threshold from the ground-truth reference audio time t_i. Commonly used thresholds range from 50 ms to 1 second depending on the application. For an automatic accompaniment system, a deviation within 50 ms would be required, while for an automatic page turner, 1 second would be still fine. We use 50 ms as the threshold in the experiments.
2) *Average Onset Deviation (AOD):* It calculates the average absolute time deviation of the alignment of all unique onsets of a piece, i.e., $\frac{1}{N_{\text{onset}}}\sum_i \left|\hat{t}_i - t_i\right|$.

Since our proposed sustained-sound reduction operation is built upon an online onset detection module, we also evaluate onset detection performance and examine its effect on score following. Again, we only consider unique onsets in the ground-

[2]http://www.openairlib.net/auralizationdb/content/lady-chapel-st-albans-cathedral. Downloaded in December, 2015.

truth. In the audio, onsets of simultaneous notes may deviate from each other slightly, but our algorithm usually only detects one onset thanks to the time threshold β. A detected onset is considered correct if it deviates from a ground-truth onset less than 30 ms. Each ground-truth onset can only be associated with at most one correctly detected onset. We use precision $P = Corr/Est$, recall $R = Corr/Ref$, and F-measure $F = 2PR/(P+R)$ to evaluate onset detection results for each piece, where $Corr$ is the number of correctly detected onsets, Est is the number of all estimated onsets, and Ref is the number of all reference onsets.

3) Parameter Settings: In all the experiments, we set frame size to 46.4 ms and hop size to 10 ms for STFT in audio spectrogram calculation. For onset detection, we set the amplitude threshold α to 40, the time threshold β to 3 frames (30 ms), and the window length threshold ω to 5 frames (50 ms). Section V-C1 analyzes how onset detection errors affect the score following results. For sustained-sound reduction, the reference frame was set to 5 frames (50 ms) before each onset frame to leave some room to cope with the onset location estimation inaccuracies. The length of the operation region L where the proposed sustained-sound reduction operation performs is set 15 frames (150 ms) immediately after each onset frame, or until the next onset if it is within the 15 frames. This parameter setting is informed by Fig. 11, and how its value affects the score following performance is analyzed in Section V-C2. Parameters of the other parts of the score following system remain the same as those reported in [12]. Note that all the parameters in this paper were set the same for all test pieces and in both non-reverberant and reverberant cases.

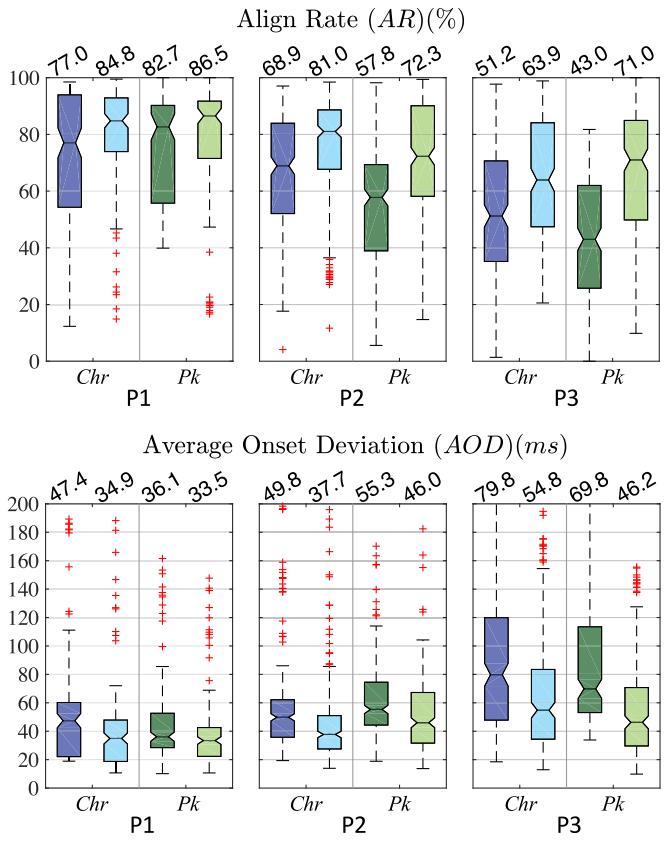


Fig. 12. Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*). The number above each box shows the median.

B. Results in Non-Reverberant Cases

We first evaluate the system on the three groups in non-reverberant cases, namely *P1*, *P2*, and *P3*. Fig. 12 shows box plots of the two score following measures (AR and AOD) of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (blue) and the spectral-peak representation (green). Due to the randomness of particle filtering in the score following framework, we ran each system 10 times per piece. Therefore, each box (along with outliers shown as red crosses) represents 10 $\times$ #Pieces data points, e.g., 120 points for each box in the group *P1*. The median of each box is shown on the top. To avoid clutter, we cut off some outliers in the figures of AOD.

We conducted paired sign tests on AR between the baseline and proposed system using both audio representations and in the three groups (6 pairs), and all the improvements passed the significance level of 0.005. More interesting observations can be made from Fig. 12. First, comparing the three recording groups, we see that score following performance using both representations generally degrades when the sustained effect becomes stronger (*P1*→*P2*→*P3*) for both measures. This degradation is especially pronounced for the baseline system, while is slighter for the proposed system. Given that the piece complexity of the three groups are similar, this shows that the sustained effect is indeed an issue for piano score following and the proposed approach is able to alleviate this issue. Second, as the sustained effect becomes stronger, the improvement of our proposed system also grows, e.g., 3.8% in *P1* and 28% in *P3*, using the spectral-peak representation. Third, comparing the two audio representations, the spectral-peak representation yields more pronounced improvement in the proposed system over the baseline system. This is because the peak removal operation in this representation removes the sustained peaks entirely, while in the chromagram representation the sustained sound is only reduced (see Eq. (5)). However, the entire removal makes the system sensitive to onset detection errors, as analyzed in Section IV-C3 and shown in the following experiments.

As online onset detection is an important module of the proposed approach, we also evaluate its performance. Fig. 13 shows the average precision, recall, and F-measure curves of onset detection within each recording group by varying the amplitude threshold α from 1 to 80 with a step of 1. It can be seen that as the sustained effect becomes stronger, onset detection performance becomes worse. The best F-measure is achieved around $\alpha = 35$ for all groups. However, the finally chosen threshold is $\alpha = 40$, as shown by the dashed vertical lines, to prefer a higher precision than recall, as explained in Section IV-C3. Note that we use the same threshold throughout all experiments, although the optimal threshold differs for different pieces and groups. For the second threshold α' used in spectral subtraction for the

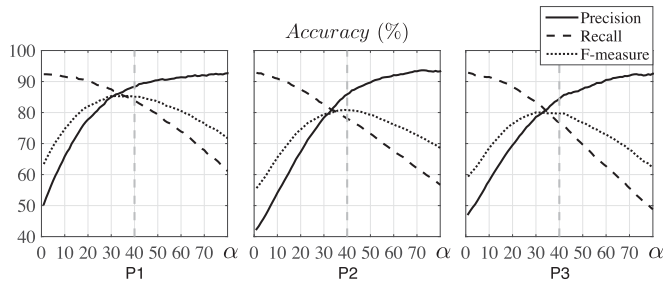


Fig. 13. Onset detection results averaged over pieces within each recording group. Note that the three curves do not necessarily intersect at the same point because they are average values.

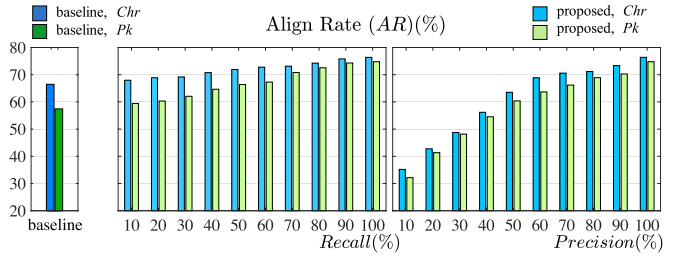


Fig. 15. Score following performances (average of AR) of the baseline system and the proposed system built on artificially controlled onset detection results on *P2*. Precision is fixed at 100% in the middle panel while recall is fixed at 100% in the right panel.

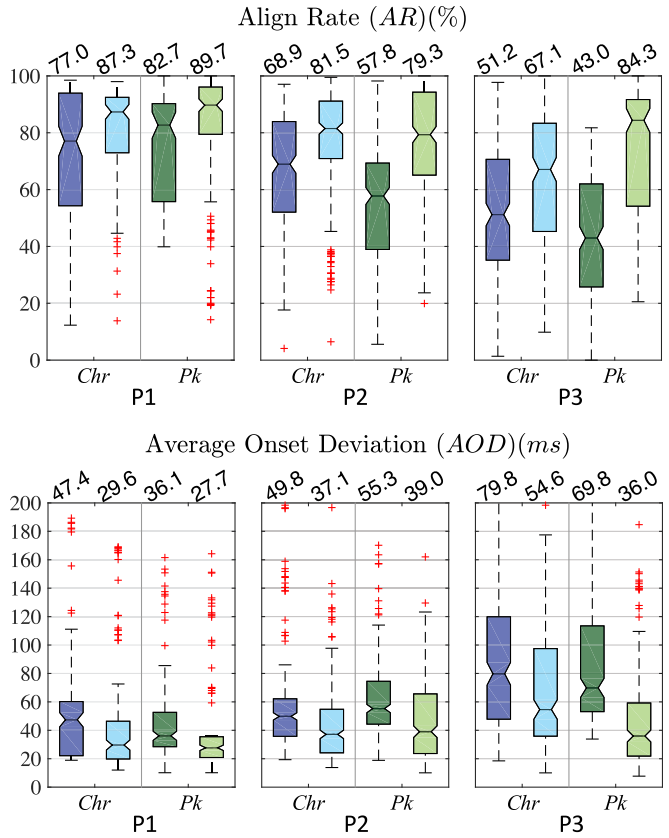


Fig. 14. Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when the ground-truth onsets are used. The number above each box shows the median.

chromagram representation, we set it to 70, which leads to a high precision for most pieces and an acceptable recall.

To isolate the effects of onset detection on the final score following performance, we also evaluate the proposed system with ground-truth onset information. Fig. 14 shows results. We can see that the proposed system shows more observable improvement over the baseline system. The improvement is significant in all cases under a sign test at a significance level of 10^{-3}. It reaches to a median AR of 89.7%, 79.3%, and 84.3% for the three recording groups, respectively, using the spectral-peak representation. Also note that *P3* with heavy sustained effect reaches a comparable value with the other groups. These values set the upper bound for the proposed approach when more advanced onset detection technique is employed. Another interesting observation is that the spectral-peak representation yields better performance than the chromagram representation in this case, while their performances are similar in Fig. 12. This is because the "bolder" peak removal operation in the spectral-peak representation, which is sensitive to onset detection false positives as analyzed in Section IV-C3, removes the sustained effect more effectively than the "more conservative" spectral subtraction operation in the chromagram representation.

C. Parameter Analysis

1) Effects of Onset Detection Errors: As analyzed in Section IV-C3, miss errors and false positive errors of onset detection have different effects on score following in our proposed approach. To further investigate this, we artificially created onset detection results with separate controls of precision and recall. Starting from the ground-truth onsets, we randomly removed some onsets to control the recall while maintaining the precision at 100%. We also randomly added false positive onsets to control the precision while maintaining the recall at 100%. Taking the group *P2* as an example, score following performance (average of AR values for all the pieces) based on these onset detection results are shown in Fig. 15. We can see that the result is indeed less sensitive to the recall (miss errors) when the precision is 100% (middle panel), while it is sensitive to precision (false positive errors) even when the recall is 100% (right panel). When the precision is less than 50%, the proposed approach shows inferior performance than the baseline system. Note that 0% recall in the middle panel (i.e., no onset is detected) would make our proposed approach the same as the baseline system.

In Fig. 16, we further show the boxplots of AOD with recall fixed as 100% and precision varing for both chromagram and spectral-peak representations. The vertical axis covers a wide range to show outliers while the numbers above the figures show medians. We can see that lowering precision causes significant increase of the median AOD value and the number of outliers with extremely large deviations (e.g., >1000 ms). In fact, those outliers correspond to runs when the system was totally lost during score following. This usually happens on pieces with sparse notes. Comparing the two audio representations, *Pk* shows more

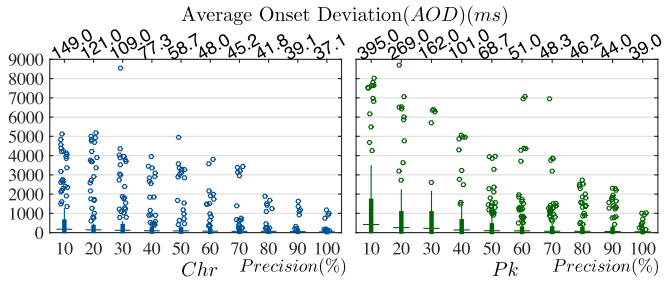


Fig. 16. Boxplot of the Average Onset Deviation of the proposed approach on artificially controlled onset detection results on the group *P2*, where recall is fixed at 100% and precision is varied. Numbers above show the medians.

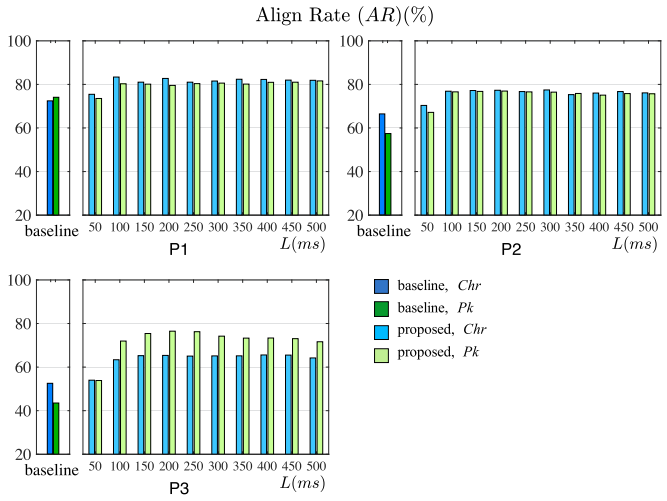


Fig. 17. Score following performance (average of AR) of the baseline system and the proposed system with different lengths of the sustained-sound reduction operation region for the three groups.

outliers than *Chr*, especially when precision is low. This supports our claim in Section IV-C3 that low precision is more harmful for the spectral-peak representation.

2) Effects of the Operation Region Length: To investigate the sensitivity of the proposed system on the sustained-effect operation length L, we conducted another experiment on the three groups in the non-reverberant case with different values of the parameter. Fig. 17 shows the results. We can see that the average align rate for P1 and P2 stays stable once L reaches 100 ms. The average value of Align Rate for P3 achieves the highest value when L reaches 150 ms and then stays stable with the chromagram representation but slightly decreases when $L > 250$ ms with the spectral-peak representation. This result shows a good correspondence with the statistics of the sustained-effect region length (SRL) in Fig. 11, where the average SRL for the three groups are 55.5 ms, 118 ms and 168 ms, respectively. When a piece has a larger SRL, the operation region should be longer as well. Interestingly, the experiment also shows that there is almost no negative effect if L is set too long. This again validates our analysis in Section IV-C1 and Fig. 6 that the operation is negligible when it is wrongly applied to regions with little or no sustained sound.

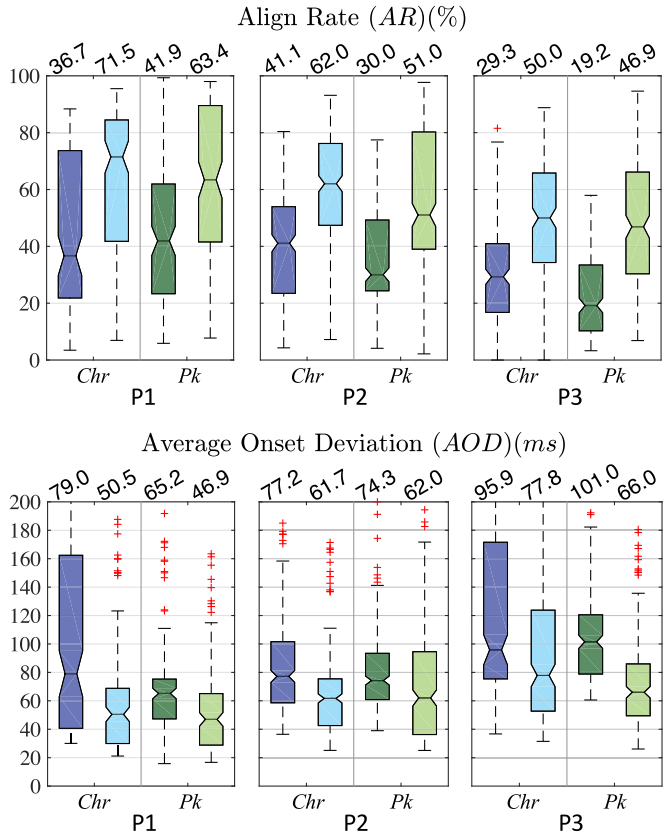


Fig. 18. Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when reverberation is added to the audio. The number above each box shows the median.

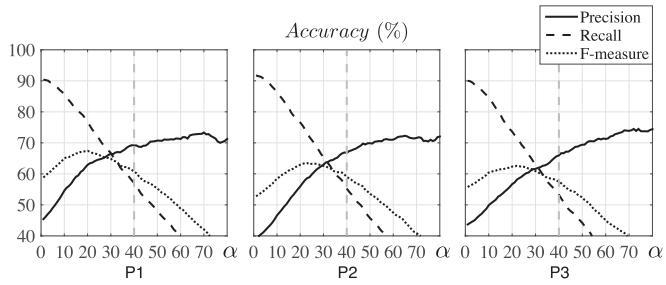


Fig. 19. Onset detection results averaged over pieces within each recording group with reverberation added. Note that the three curves do not necessarily intersect at the same point because they are average values.

D. Results in Reverberant Environments

In this section, we further evaluate the proposed approach in reverberant environments. Reverberation imposes the sustained effect on all notes of a piece. It also blurs the onsets and makes onset detection challenging. As described in Section V-A1, we impose reverberation on all the 60 pieces in our dataset. Figs. 18 and 19 show the score following and onset detection result respectively. Comparing with Fig. 13, we can see that the performance of our online onset detection degrades from anechoic environments to reverberant environments. For example, the best F-measure drops from 81% to

64% for the group *P2*. The threshold α that achieves the highest F-measure is now around 20. However, we still chose the same threshold 40 in the experiment to prefer higher precision than recall.

Fig. 18 shows the score following results. Comparing with Fig. 12, we can see that both the baseline system and the proposed system degrade dramatically. However, the improvement of the proposed system over the baseline system becomes more pronounced, especially in the *P1* group, as they now have the sustained effect due to reverberation. The improvement on AR is significant in all settings under a sign test at the significance level of 10^{-14}. Another interesting observation is that the baseline system performs badly in some pieces from group *P1*. Further investigation indicates that some pieces are performed with fast-running notes (see Fig. 10(b)). We argue that the reverberation blurred and blended these notes so much that the baseline system was not able to follow the pieces and result in a large number of outliers with extremely large deviations (not shown completely in the figures). The proposed system, however, was able to greatly alleviate this issue through the sustained-sound reduction operation for both chromagram and spectral-peak representations.

Audio examples of the above experiments can be accessed at http://www.ece.rochester.edu/∼bli23/projects/pianofollowing.html.

VI. Conclusions

In this paper we proposed a score following approach to follow piano audio performances with the sustained effect. The sustained effect refers to the phenomenon that notes are sustained longer than their notated length in the score. It blends the sustained notes with new notes, introduces audio-score mismatch, and causes delay errors of score followers. We analyzed three main causes of the sustained effect, namely the legato playing technique, the usage of the sustain and sostenuto pedals, and the room reverberations. We proposed an approach to reduce the sustained effect by reducing the sustained spectral components in a number of frames immediately after each detected onset. This approach was developed for two commonly used audio representations, the chromagram and the spectral-peak representations. We integrated the proposed approach with a Markov model-based score following framework to test its effectiveness. We also analyzed effects of the proposed approach when it is wrongly applied. Systematic experiments showed that the proposed approach improved the score following accuracy and robustness significantly on a variety of piano pieces with different degrees of the sustained effect. Detailed analysis of the rationales of parameter settings and their effects on score following were also provided.

For future work, we plan to consider other specific properties of piano music to improve the alignment performance. For example, the sound decay may result in potential mismatch around note offset frames. In this case, a time-varying matching function that considers the exponential energy decay would improve the alignment accuracy.

Acknowledgment

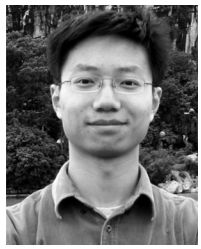
The authors would like to thank the three anonymous reviews for their insightful and constructive comments, which have greatly improved this article.

References

[1] M. Puckette and C. Lippe, "Score following in practice," in *Proc. Int. Comput. Music Conf.*, 1992, p. 182.
[2] V. Thomas, C. Fremerey, D. Damm, and M. Clausen, "Slave: A score-lyrics-audio-video-explorer," in *Proc. Int. Soc. Music Inf. Retrieval*, 2009, pp. 717–722.
[3] Q. Wang, Z. Guo, G. Liu, C. Li, and J. Guo, "Local alignment for query by humming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3711–3715.
[4] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed transcription for automatic piano tutoring," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 2153–2157.
[5] C. Raphael, "Music plus one and machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 21–28.
[6] Tonara. (2015). [Online]. Available: http://www.tonara.com
[7] M. Prockup, D. Grunberg, A. Hrybyk, and Y. E. Kim, "Orchestral performance companion: Using real-time audio to score alignment," *IEEE Multimedia*, vol. 20, no. 2, pp. 52–60, Apr.–Jun. 2013.
[8] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, 2011.
[9] T. Itohara, K. Nakadai, T. Ogata, and H. Okuno, "Improvement of audio-visual score following in robot ensemble with human guitarist," in *Proc. 12th IEEE-RAS Int. Conf. Humanoid Robots*, 2012, pp. 574–579.
[10] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proc. Eur. Conf. Artif. Intell.*, 2008, pp. 241–245.
[11] S. Gordon, *A History of Keyboard Literature: Music for the Piano and its Forerunners*. Belmont, CA, USA: Wadsworth, 1996.
[12] Z. Duan and B. Pardo, "A state space model for online polyphonic audio-score alignment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 197–200.
[13] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.
[14] B. Li and Z. Duan, "Score following for piano performances with sustain-pedal effects," in *Proc. Int. Soc. Music Inf. Retrieval*, 2015, pp. 469–475.
[15] M. Puckette, "Score following using the sung voice," in *Proc. Int. Comput. Music Conf.*, 1995, pp. 175–178.
[16] L. Grubb and R. Dannenberg, "A stochastic method of tracking a vocal performer," in *Proc. Int. Comput. Music Conf.*, 1997, pp. 301–308.
[17] N. Orio and F. Déchelle, "Score following using spectral analysis and hidden Markov models," in *Proc. Int. Comput. Music Conf.*, 2001, pp. 151–154.
[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] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization." in *Proc. Int. Soc. Music Inf. Retrieval*, 2006, pp. 192–197.
[20] H. Kaprykowsky and X. Rodet, "Globally optimal short-time dynamic time warping, application to score to audio alignment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 249–252.
[21] S. Dixon, "Live tracking of musical performances using on-line time warping," in *Proc. Int. Conf. Digital Audio Effects*, 2005, pp. 92–97.
[22] A. Arzt and G. Widmer, "Simple tempo models for real-time music tracking," in *Proc. Sound Music Comput. Conf.*, 2010.
[23] B. Pardo and W. Birmingham, "Modeling form for on-line following of musical performances," in *Proc. 20th Nat. Conf. Artif. Intell.*, 2005, pp. 1018–1023.
[24] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 185–188.
[25] D. P. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. IV-1429–IV-1432.

[26] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2009, pp. 1869–1872.
[27] A. Arzt, G. Widmer, and S. Dixon, "Adaptive distance normalization for real-time music tracking," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 2689–2693.
[28] S. Wang, S. Ewert, and S. Dixon, "Robust joint alignment of multiple versions of a piece of music," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 83–88.
[29] M. Müller and S. Ewert, "Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proc. Int. Conf. Music Inf. Retrieval*, 2011, pp. 215–220.
[30] S. Dixon and G. Widmer, "MATCH: A music alignment tool chest," in *Proc. Int. Soc. Music Inf. Retrieval*, 2005, pp. 492–497.
[31] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.
[32] N. Montecchio and N. Orio, "A discrete filter bank approach to audio to score matching for polyphonic music," in *Proc. Int. Soc. Music Inf. Retrieval*, 2009, pp. 495–500.
[33] A. Cont, "Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 245–248.
[34] J. Carabias-Orti, F. Rodriguez-Serrano, P. Vera-Candeas, N. Ruiz-Reyes, and F. Canadas-Quesada, "An audio to score alignment framework using spectral factorization and dynamic time warping," in *Proc. Int. Soc. Music Inf. Retrieval*, 2015, pp. 742–748.
[35] C. Joder and B. 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.
[36] 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. 51–59.
[37] B. Niedermayer, S. Böck, and G. Widmer, "On the importance of "real" audio data for MIR algorithm evaluation at the note-level - a comparative study," in *Proc. Int. Soc. Music Inf. Retrieval*, 2011, pp. 543–548.
[38] R. Bresin and G. Umberto Battel, "Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the andante movement of Mozart's Sonata in G major (K 545)," *J. New Music Res.*, vol. 29, no. 3, pp. 211–224, 2000.
[39] B. H. Repp, "Acoustics, perception, and production of legato articulation on a digital piano," *J. Acoust. Soc. Amer.*, vol. 97, no. 6, pp. 3862–3874, 1995.
[40] H. M. Lehtonen, H. Penttinen, J. Rauhala, and V. Välimäki, "Analysis and modeling of piano sustain-pedal effects," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1787–1797, 2007.
[41] Reverberation time. (2015). [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/acoustic/revtim.html#c1
[42] J. Meyer, *Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers*. New York, NY, USA: Springer, 2009.
[43] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.
[44] S. Dixon, "Onset detection revisited," in *Proc. Int. Conf. Digital Audio Effects*, 2006, pp. 133–137.
[45] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in *Proc. Int. Soc. Music Inf. Retrieval*, 2010, pp. 589–594.
[46] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proc. Int. Soc. Music Inf. Retrieval*, 2012, pp. 49–54.
[47] X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in *Proc. Int. Comput. Music Conf.*, 2001, pp. 30–33.
[48] T. Fujishima, "Realtime chord recognition of musical sound: A system using Common Lisp Music," in *Proc. Int. Comput. Music Conf.*, 1999, pp. 464–467.
[49] Chroma implementation. (2015). [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/chroma-ansyn/
[50] W. Goebl, "Melody lead in piano performance: Expressive device or artifact?" *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 563–572, 2001.
[51] *Java MIDI tools*. (2016). [Online]. Available: http://www.ee.columbia.edu/csmit/java_midi.html
[52] A. Cont *et al.*, "Evaluation of real-time audio-to-score alignment," in *Proc. Int. Soc. Music Inf. Retrieval*, 2007, pp. 315–316.

Bochen Li received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2014. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA. His research interests primarily include the interdisciplinary area of signal processing and machine learning toward music information retrieval applications based on audio-visual multimodal analysis, such as video-informed multipitch estimation and streaming, source separation, and visual association.



Zhiyao Duan (S'09–M'13) received the B.S. and M.S. degrees in automation from Tsinghua University, Beijing, China, in 2004 and 2008, respectively, and the Ph.D. degree in computer science from Northwestern University, Evanston, IL, USA, in 2013. He is an Assistant Professor with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA. His research interests include the broad area of computer audition, i.e., designing computational systems that are capable of analyzing and processing sounds, including music, speech, and environmental sounds. Specific problems that he is working on include automatic music transcription, multipitch analysis, music audio-score alignment, sound source separation, and speech enhancement.