MuEns: A Multimodal Human-Machine Music Ensemble for Live Concert Performance

Akira Maezawa **Kazuhiko Yamamoto***
Yamaha Corporation
203 Matsunokijima, Iwata, Shizuoka, 438-0192, Japan
akira.maezawa@music.yamaha.com, yamo_o@acm.org

ABSTRACT
Musical ensemble between human musicians and computers is a challenging task. We achieve this with a concert-quality synchronization using machine learning. Our system recognizes the position in a given song from the human performance using the microphone and camera inputs, and responds in real-time with audio and visual feedback as a music ensemble. We address three crucial requirements in a musical ensemble system. First, our system interacts with human players through both audio and visual cues, the conventional modes of coordination for musicians. Second, our system synchronizes with human performances while retaining its intended musical expression. Third, our system prevents failures during a concert due to bad tracking, by displaying an internal confidence measure and allowing a backstage human operator to "intervene" if the system is unconfident. We show the feasibility of the system with several experiments, including a professional concert.

ACM Classification Keywords
H.5.1. Multimedia Information Systems: Audio input/output; H.5.5. Sound and Music Computing: Signal analysis, synthesis, and processing; J.5. Arts and Humanities: Music

Author Keywords
Human-machine Music Ensemble; Multimodal Interaction; Machine Learning; Score Following; Live Concert System

INTRODUCTION
When human musicians perform in a music ensemble, they interact with each other to enrich the musical expression. During the performance, they would dynamically change the tempo and the articulations on the spot, coordinating with each other through eye movements, various gestures, and auditory cues. To make computers to imitate such interactions has been desired for decades. We show a practical concert-quality system for achieving this challenge. Our goal in this study is to realize a machine with the capability to coordinate timing with human musicians, as a human musician would.

A possible approach is automatic accompaniment using score following [10]. Score following technique recognizes the corresponding position in a given song from the audio of human performances, and plays back the machine part (*e.g.*, the accompaniment part). However, through a preliminary user study, we found this technique has three serious problems for coordinating the timing between a machine music sequencer and human musicians in real-time. First, it requires the recognition and generation of auditory and visual cues, but incorporating such multi-modal cues is difficult by existing works. For example, musicians nod to each other to anticipate the timing of the beginning of a piece; at the same time, once the music starts, they would listen to each other to grasp the ebb and flow of tempo. Second, the machine should synchronize to human musicians while retaining its expression of tempo, but the balance between these conflicting goals change dynamically during the piece. For example, when the machine sequencer is playing the melody, the human players would follow the machine sequencer and appreciate its sequenced nuances. Conversely, when the humans are playing the melody, the human players would prefer the machine to follow them while retaining its nuance. Finally, the system must never fail during a concert, but such a guarantee is difficult to make with a fully automatic coordination system. That is, the tracking of the subtle multi-modal cues in music ensemble is a difficult task that is prone to errors; yet, if the machine responds highly incorrectly, the concert would be ruined, which is fatal as a professional staged event.

To address these problems, we propose a novel music ensemble system, *MuEns*, that enables a multi-modal, flexible and error-robust music ensemble between human musicians and a machine (Figure 1). Our system integrates the auditory and the visual cues of human performers, and reacts to them with an automatic playback of a pre-recorded music data (in standard Musical Instruments Digital Interface (MIDI) file) and a pre-choreographed visual feedback. It also balances between synchronizing to the human musicians and retaining the musical nuance of the pre-recorded music data. To avoid the potentially critical errors of the automatic tracking algorithm during a live concert, our system also allows a human operator to intervene on-the-fly and guide the system.

We describe how our system was improved through an iterative design process. Our first trial was a pilot study with a

*Also affiliated with the University of Tokyo

CHI'17, May 06 – 11, 2017, Denver, CO, USA.

ACM 978-1-4503-4655-9/17/05...$15.00.
DOI: http://dx.doi.org/10.1145/3025453.3025505

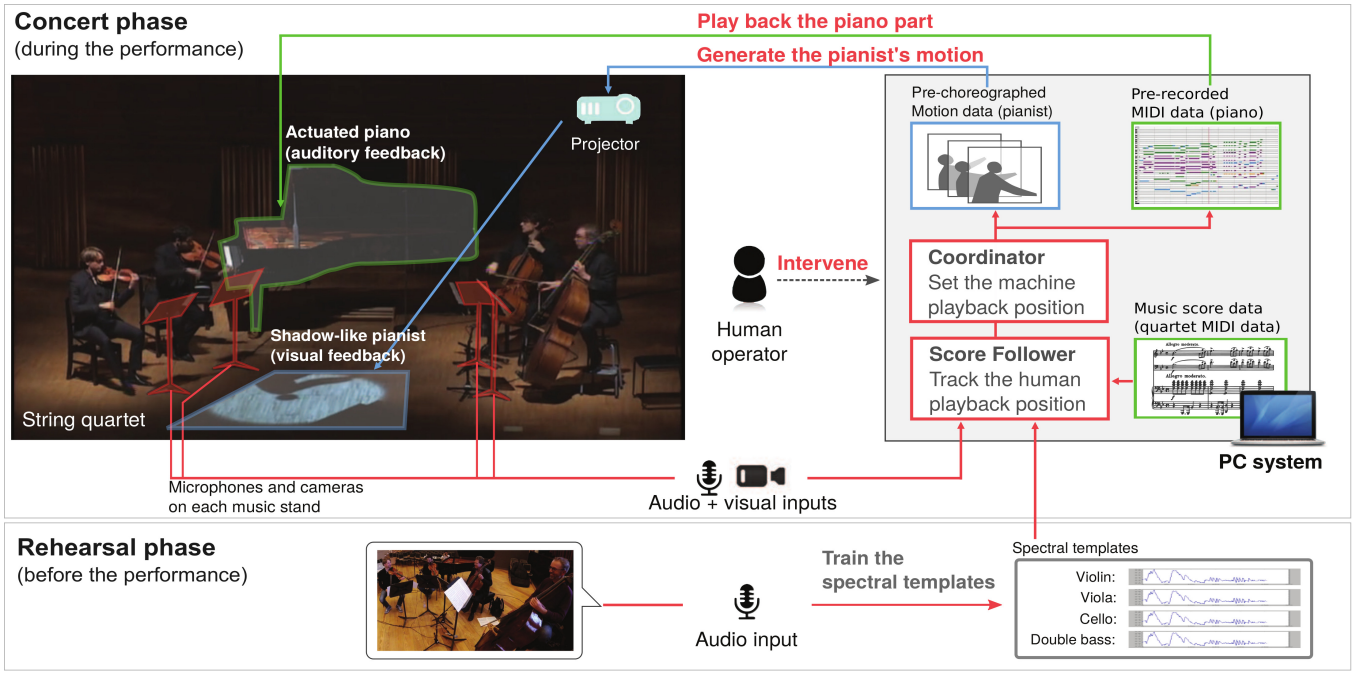

Figure 1. The system overview. Our system takes as inputs the audio signals from the microphone and movies from the camera of each human player. The system responds to human performances using machine learning-based tracking algorithm, and outputs (1) pre-recorded MIDI data that drives the actuated piano and (2) pre-choreographed visual feedback data that represents the behavior of the system. The training data, required for tracking the playback position of the stage performance, is recorded during the rehearsal. To avoid fatal errors, a human operator observes the confidence of the system and manually intervenes as necessary.

preliminary design of the system in preparation. To overcome the problems encountered in this pilot study, we designed the final system and had a professional concert to assess its feasibilities. Finally, we show the results with the reception of the audience, and summarizing the rehearsals and the interviews.

Our contributions are as follows:

1. We present a novel multi-modal music ensemble system that recognizes and generates both auditory and visual cues, enabling a tight coordination between human musicians and a machine.

2. We present a musical coordination algorithm that enables to adjust the musicality and synchronicity of the machine during the performance, where the adjustment is bootstrapped automatically and refined through manual annotations.

3. We present a mechanism for preventing serious errors during a concert, by allowing an optional human operator to manually intervene with the system during the concert.

4. We show, with several experiments including a professional concert, a concert-quality system for music ensemble between the humans and a computer with machine learning.

RELATED WORK

Score Following: Score following is the technique to continuously track the position of human performance in a given music score, and is a critical component in automatic music accompaniment. The key point of a score follower is to jointly express (1) the similarity between the current observation and the expected observation in each position in the music score and (2) the allowed temporal evolution of the score position. This is achieved through probabilistic inference [14, 28, 6] or path optimization [8, 15, 2]. Furthermore, since the temporal evolution is better captured by explicitly expressing the transition of the underlying tempo, the underlying tempo curve is often inferred as well [28, 35, 6, 18, 23, 24].

Current score following systems do not recognize visual cues, which human musicians use for coordinating parts of the music where no one is playing [31]. For example, nodding gestures are often used to synchronize the start of a song. While some studies do use visual information, such as the periodic hand motion of a guitarist [16], they are intended to improve the tracking of what is already trackable using audio signal alone; yet, visual cue is critical when the audio signal alone is insufficient for tracking the human musicians. The importance of cues in a music ensemble system has been pointed out [13], but has not been applied for automatic accompaniment using score following. In a similar problem of beat tracking, visual information has been incorporated to aid tracking [21, 4].

Automatic Music Accompaniment: There are some systems that use score following to create a machine accompaniment that synchronizes to human players [29, 5]. There are three important issues in automatic accompaniment. First, the accompaniment should not only synchronize to humans, but also sound musical. In other words, an ensemble system should encode the musicality of the machine part and encode how it should coordinate. Second, the musicality of the machine part

should be adjustable: performers may want to change the coordination strategy or nuances, so the system should respond to these requests, or at least be adjustable by a human operator. Third, to use the system in a professional concert, a scheme to guarantee a failure-free concert is required [26] – the show must go on even if the score follower makes mistakes.

To achieve these goals, Antescofo system [5] allows programming of electronic events that are synchronized to musicians, and allows the music composer to choose what kind of synchronization is required. However, it is incapable of balancing the musicality of the machine part and synchronization. Safety measures against failed tracking is achieved through the choice of a conservative synchronization algorithm with a limited flexibility to follow the musicians. Such a scheme, however, requires the composer to anticipate everything that might go wrong on stage, that might adversely affect the score follower; this is a difficult task for the composer.

Music-plus-one system [29] uses a dynamic Bayesian network to learn the temporal pattern of the machine and the human players [27]. It is, however, incapable of directly describing how the machine and the human players should coordinate. Safety measures against failed tracking is achieved by reporting the output of the score follower only when its confidence has increased significantly. This kind of approach, however, is incapable of handling mistakes where the score follower confidently reports wrong positions; this kind of error occurs, for example, a piece contains many short repeated segments.

To jointly learn the coordination strategy and musicality, elaborate machine learning approaches have been proposed [32, 33], but they require multiple annotated instances of ensemble among humans for each piece to play, making it costly to prepare a working system.

Other Interactive Music Systems: Instead of automatic accompaniment, the accompaniment may be controlled manually, for example, through tapping of the beats [9] or conducting [1]. In these systems, the musicality between the human and the machine is balanced by adjusting the tempo based on the human input and a template tempo curve. Such a manually controlled system is robust because the tracking of the human musicians is delegated to a human operator. It however needs a human to control the system during the entire performance. It thus requires high musical skills for the operator, as the operator participates as a musician in the ensemble.

Furthermore, in one extreme of human-machine music ensemble, a human could play along with an accompaniment data, *a la* karaoke. In this case, tighter coordination is possible if the human could anticipate the machine behavior through a visual feedback [34]. Visual feedback allows the user to anticipate the upcoming notes played by the system. However, since the system is incapable of following the user, it is impossible for the user to express music through tempo changes. At the other extreme, the machine could generate a completely improvisatory response to human playing [25], but such an approach is inapplicable if the machine should play back a specific accompaniment in sync with the user.

PRELIMINARY USER STUDY

To design the system, we have collaborated with four top-level string instrument players (violin, viola, cello, and double bass) from the Scharoun Ensemble of the Berliner Philharmoniker, and have conducted the design process iteratively together for two months. We used the fourth and the fifth movements of the "Trout" quintet by F. Schubert for the experimental song. The piano part was played by a Yamaha Disklavier player piano, driven by the proposed system. The Disklavier player piano is an acoustic piano that can be driven from an external device using the MIDI protocol. The MIDI sequence data of the player piano was created by a professional pianist who recuperated a recording of the "Trout" quintet from 1980, played by Sviaslatov Richter, a legendary Russian pianist of the 20th century, and the Borodin Quartet. The performers were told that this is an experiment towards a concert featuring an ensemble between legendary musicians of the 20th century and the 21st century, bridged using computer technology.

We first investigated how human players perform with each other and with a computer using the preliminary system. Upon the pilot study, we designed our final system and had a professional concert to validate its feasibilities.

Design Strategy and Implementation

To investigate how human musicians play with each other, we begin with three assumptions for music performance: (1) auditory information is sufficient to enable the coordination between the musicians, (2) the sequence of playing speed at each position in the song (tempo curve) of the human performers would fluctuate about a default tempo curve with a constant variance throughout the piece, and (3) the extent to which the system should fix its timing is fixed throughout the piece. Under these assumptions, the system was designed such that it tracks the fluctuation of tempo about a fixed trajectory, using the audio signal of human musicians' playing.

Our first prototype system is shown in Figure 2. The system takes audio data of each player from microphones as input, and follows them with playback of a pre-recorded MIDI data by a player piano. The system consists of two sub-systems: Score follower and Coordinator. The score follower estimates the corresponding position from the microphone input using a hidden Markov Model (HMM). Using this timing information emitted from the score follower, the coordinator estimates the playback position of the pre-recorded MIDI data and sends necessary MIDI messages to the player piano.

Modeling the Human Performance

We use a hierarchical HMM approach for the score follower. In this HMM, a time series of the constant-Q transform (CQT) and ΔCQT from the microphones becomes the observation, and the corresponding position in a given musical score becomes the hidden state. The hidden state is described hierarchically. Namely, it divides the song into multiple segments, and each segment consists of multiple left-to-right Markov models, each n of which subdivides the segment with a different resolution of the score position and assigns a large probability for the transition to the next position in the score. Thus, each state inside each segment is associated with (1) the position

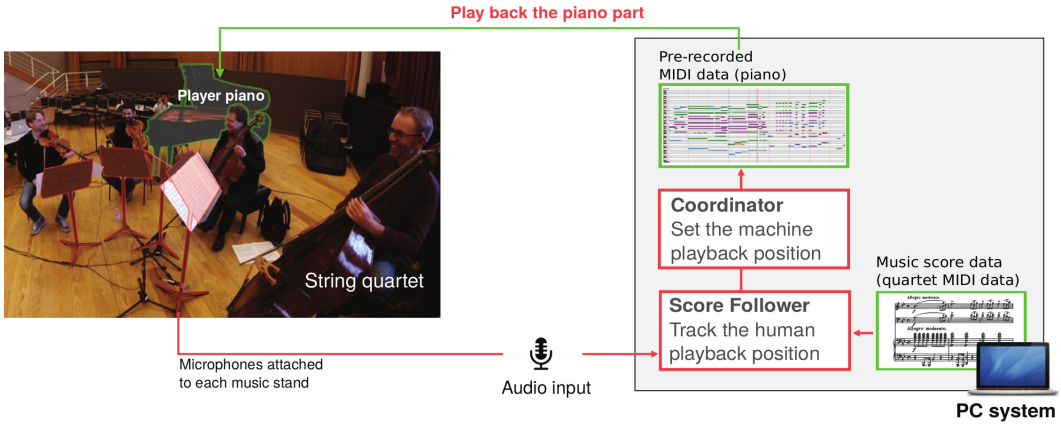

Figure 2. The preliminary design. The system takes audio data of each player from microphones as input, and follows them with playback of a pre-recorded MIDI data by a player piano.

in the given song, (2) the expected observation, and (3) the expected number of frames that is required to play the frame. This means the selection of n enables the system to roughly decide the tempo, and the transition probability adjusts for any mismatch between the expected tempo and the actual playing. Specifically, if we let r be the segment index, n be the expected duration inside the segment and l be the elapsed frame within the segment, we express the state transition in terms of three cases: (1) state (r,n,l) transitions to itself with probability $2p$, (2) state $(r,n,l<n)$ transitions to state $(r,n,l+1)$ with probability $1-p$, and (3) state $(r,n,n-1)$ transitions to $(r+1,n',0)$ with probability $(1-p)0.5\lambda^{-1}\exp(-\lambda|n'-n|)$ for some $\lambda>0$. For each position in the score s (which is associated with one or more state of the HMM (r,n,l)), we express the likelihood of observing the CQT $\mathbf{c}\in\mathbb{R}^F$ and the ΔCQT $\mathbf{d}\in\mathbb{R}^F$ as follows:

$$\log p(\mathbf{c},\mathbf{d}|s,\kappa_s,\eta_s,\bar{\mathbf{c}}_s,\bar{\mathbf{d}}_s) = \text{const.} + \kappa_s \mathbf{c}^T\bar{\mathbf{c}}_s + \eta_s \mathbf{d}^T\bar{\mathbf{d}}_s, \quad (1)$$

where $\bar{\mathbf{c}}_s$ and $\bar{\mathbf{d}}_s$ are the representative CQT and ΔCQT at position s, and κ_s and η_s are scaling parameters. The expected values of $\bar{\mathbf{c}}$ and $\bar{\mathbf{d}}$ are extracted from the music score. For each pair of pitch and instrument i, there is an associated "template" CQT $\mathbf{w}_i$. Then, given h_{si}, the loudness of pair i at score position s, the representative CQT is given as follows:

$$\bar{\mathbf{c}}_s = \sum_i h_{s,i}\mathbf{w}_i. \quad (2)$$

$\bar{\mathbf{d}}$ is obtained by taking the adjacent difference of $\bar{c}_s$ over s and half-wave rectifying it. The system, after a rehearsal, uses the history of the tracked position and the recorded audio to update $\mathbf{w}_i$ using Poisson-Gamma nonnegative matrix factorization [3]. Here, the prior distribution is based on the tracked position and the expected notes at each position, similar to [11].

When the nth new note is played at time τ_n, the estimated playback position μ_n and the variance of the estimate σ_n^2 are emitted from the score follower. These estimates are obtained by evaluating the Laplace approximation of the posterior distribution over the score position. The coordinator receives (τ_n,μ_n,σ_n) and predicts the playback position of the machine part at time t, $m(t)$. The coordinator assumes that $m(t)$ is played by a piecewise-constant velocity v_n (playback position of the score [second]-per-second) with offset x_n (playback position [second]), based on the most recent information received from the score follower:

$$m(t) = x_n + v_n(L+t-\tau_n). \quad (3)$$

Here, L is the input-to-output latency of the system (about 300 milliseconds). The coordinator infers v_n and x_n by assuming an underlying process given as follows:

$$\begin{aligned} x_n &= x_{n-1} + \Delta T_{n,n-1}v_{n-1} + \varepsilon_{n,0} \\ v_n &= \beta\bar{v}_n + (1-\beta)v_{n-1} + \varepsilon_{n,1} \quad (4) \\ \mu_n &\sim \mathcal{N}(x_n,\sigma_n^2), \quad (5) \end{aligned}$$

where $\Delta T_{n,m} = \tau_n - \tau_m$ and $\mathcal{N}(\mu,\sigma^2)$ is a Normal distribution with mean μ and variance σ^2. ε is a zero-mean Gaussian noise, the variance of which governs the extent to which the machine synchronizes to the human players, and how much the machine part's tempo may fluctuate based on our assumption. This model assumes that human players play a given piece of music with more-or-less a similar tempo trajectory, an useful assumption in modeling timing across different interpretations to a same piece of piece [22]. $\bar{v}_n$ is the "default" velocity at position x_n, obtained by analyzing the velocity trajectory (*i.e.*, the tempo curve) of a music performance of the same piece by a human ensemble. β is a fixed scalar that determines how strongly v_n reverts to the default tempo. We call this β the *mean-reverting* parameter.

Preliminary Experiment

Before using the system with the ensemble, we preliminarily evaluated the system's components.

First, the score follower was evaluated quantitatively using an in-house piano performance dataset. The dataset consists of ten piano etudes by Burgmuller played by a professional pianist, and corresponding MIDI files for the system to follow. We evaluated the piecewise precision [7], the percentage of the note onset timings reported by the score follower that are within 300 ms of the correct onset timings. Our system obtained a piecewise precision of 96%.

Second, the coordinator was evaluated through a subjective evaluation using professional pianists. First, we prepared piano pieces with audio accompaniments. We chose pieces such that (1) a wide variety of genre was covered, from piano concerto to popular music, and (2) the score follower rarely made significant mistakes. Next, four pianists tested the system without mean reversion ($\beta = 0$) for one day, and five pianists tested the system with mean reversion ($\beta = 0.01$) for one day. Note that coordination without mean reversion amounts to machine accompaniment that simply smoothes the tempo curve of the human performer. The participants were asked to write down any issues, especially those regarding playability. After using the system, an informal interview with each participant was held, each spanning one hour. In the interview we asked the pianists to comment on the overall playability of the system. The group that tested the system without mean reversion all commented that the system is unusable because the system kept on getting slower or faster, perhaps since the machine kept responding to the tendency for a performer to lead or lag slightly. On the other hand, the group that tested the system with mean reversion did not mention this kind of behavior.

These evaluations suggest that (1) the score follower tracks the musicians adequately in most places but fails at a few spots and (2) the mean-reverting dynamics is effective over a simple smoothing of the score follower output.

Experiment 1

The ensemble rehearsed the experimental song using the system. This experiment took three hours and was split into three stages. In the first hour, only the human string players rehearsed (without use of the player piano). In the next hour, the human players rehearsed with a player piano that simply plays back the MIDI piano data, *a la* karaoke. Finally, the human players rehearsed with the player piano using the system.

The ensemble commented that karaoke is "weird" to play and inviable. Especially for the entrance timing of the piano, the ensemble noted that "if he [the piano entrance] is a tiny bit slow, it is not logical... The answer [response of the piano] doesn't make sense... because it is too late or too early."

When using the system, the ensemble commented that it was "incredibly great" that the entrance timing was timed properly. However, they noted that "he [the piano] loses his character" when playing the melody, which was unsatisfactory because they wanted the "Russian sound" of the MIDI data to be preserved. Furthermore, they were dissatisfied that the MIDI data always tracked the humans; they instead wanted the computer to understand leadership roles. They commented that "to decide who is leading... is the point about making chamber music," and that leadership is determined during rehearsals because it is "always different [with the ensemble]... [and is dependent on] personal taste," even though some parts in the music are "very clear" on leadership roles. They thought that the system tracks "wonderfully... like a first-class musician" when the piano should follow the humans, but "makes stupid mistakes" at places. During these "stupid mistakes," the ensemble could not continue on playing.

Discussion

Through this study, we had found several problems in our first design strategy.

First, the ensemble was dissatisfied that the music expression of the piano data has changed as a result of tracking humans too much. This kind of problem occurs for two reasons. First, the value of the mean-reverting parameter β that is appropriate for tracking humans is different from the value that is appropriate for retaining the original tempo curve of the piano $\bar{v}_n$. Second, the parameters of the coordinator, v_n and x_n, change only when the human players play a new note. This kind of logic ignores how the piano part has been played so far. For example, if, in a given phrase, there are many notes played by the piano, v_n should change more smoothly than if the piano is not playing. Thus, synchronization and playing of the piano part must be treated independently and be adjustable.

Second, the ensemble fell apart when the human players expected leadership roles while the machine synchronized to humans. This kind of problem occurs because the extent of synchronization, specified through ε, is fixed, but the musicians expect such an extent to change within a piece of music. Therefore, a mechanism to specify the degree of synchronization at selected segments in a piece is necessary.

Third, the coordination between the human musicians and the machine was poor when both parts played after a long pause. For example, the beginning of fifth movement requires the piano and strings to start simultaneously, but it was impossible to coordinate the starting timing from silence. This kind of problem occurs because the machine neither provides nor understands visual cues, which the human musicians use to complement audio information.

Finally, the ensemble fell apart when the score follower made mistakes. Specifically, when the score follower lost track of where the human players are playing, the system generated highly inconsistent responses. This kind of inconsistency at best confused the musicians and at worst made the ensemble unplayable. For a successful live performance, safety measures are required to guarantee that the performance will proceed, even if the score follower makes mistakes.

DESIGN OF MuEns

Based on the preliminary study, the design has been changed to that as illustrated in Figure 1. We call this system *MuEns*. The system takes audio and visual data of each player from microphones and cameras as the inputs, and follows them with the playback of (1) a pre-recorded MIDI data by a player piano and (2) a pre-choreographed motion data that expresses cueing gestures of a pianist. The system consists of the Score follower and the Coordinator to track and coordinate the playback.

This system provides three advantages over the preliminary system. First, synchronization is possible at points where auditory cues are unavailable, thanks to the integration of audio and visual informations. Second, a more fluid ensemble is realized through the modification of the underlying coordination algorithm. Third, a safety measure is provided in case the score follower fails, by allowing a human operator to intervene and take control over the system when necessary.

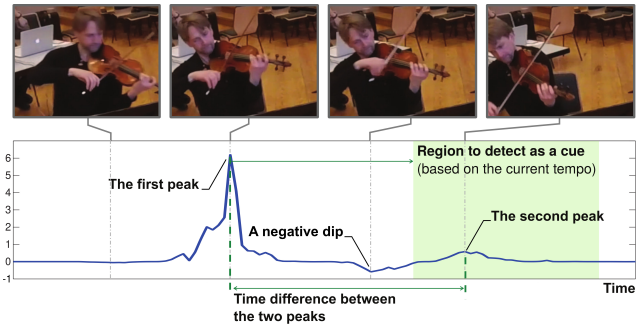

Figure 3. Detecting the visual cue motions. The blue line represents the motion feature. When a human player conducts a cue gesture, its trajectory draws two peaks beyond a threshold across a minus dip on the timeline. If the time interval between the two peaks is sufficiently close to the current beat duration, the system recognizes it as a cue.

Method

Visual Recognition and Response

The ensemble system needs a capability to both recognize and generate visual cues, depending on which instrument has the initiative at a given phrase. When musicians start a new phrase, the player who has the initiative in the new phrase would give a cue gesture at one beat before the start of a new phrase. For example, to start a piece of music, the musician who plays the melody would nod to anticipate the entry timing. Similarly, a nodding gesture is used to recover from a *fermata*, a note that is held arbitrarily long. Visual cue is important in these situations because auditory information is insufficient for timing coordination. Thus, when the human players have the initiative, the ensemble system must recognize the human players' cues to coordinate the starting timing. Conversely, when the machine part has the initiative, it should present cues to the human players for coordinating the starting timing.

To handle the case that humans have the initiative, we use visual information from the camera attached on each music stand. We call this subsystem the motion detector. The motion detector computes the time series of the average of the optical flows u_t within all the pixels from the camera images, where t denotes the frame number. We define the motion feature as the time series of the accumulated inner product $I_t = \sum_{i=t_{\text{prev}}}^{t-1} u_t \cdot u_i$, where t_{prev} determines how number of the previous frames are considered. When a human player conducts a cue gesture (Figure 3), (1) the motion feature first increases (moves a certain direction continuously), and after exceeding a threshold, (2) it decreases to minus value (turns back to the opposite direction), and (3) it again increases subsequently (back to the natural position). This trajectory draws two peaks beyond a threshold across a minus dip on the timeline. When the motion detector detects this trajectory, we compute the time difference between two peaks, and compare it to the current estimated beat duration (the inverse of the tempo) by the score follower. If both times are sufficiently close, the motion detector recognizes it as a cue motion (we assume the tempo of a song would not change rapidly), and sends the expected timing of the next beat to the score follower.

During the stage performance, the score follower receives the cue informations from the motion detector as described above. Additionally, during the rehearsals, we annotate the positions $\{\hat{q}_i\}$ in the musical score where visual cue might be expected. When the detected cue position is sufficiently close to one of the $\{\hat{q}_i\}$, the score follower sets the likelihood of the positions $\cup[\hat{q}_i - \tau, \hat{q}_i]$ for $\tau > 0$ to zero. This leads the posterior distribution to "avoid" score positions before the cue positions. Unlike the existing method based on top-down integration of cues [21], with the bottom-up integration of the visual information to the HMM-based score follower, the uncertainty of audio input can be preserved.

Figure 4. The imaginary pianist-like projection indicates the timing and the tempo to begin a new phrase, when the machine has the initiative to start the phrase.

On the other hand, when the system has the initiative for the beginning timing of the next phase, the system indicates the timing to the human players by a visualization (Figure 4), similar in spirit to [34]. This visualization is projected on the floor among the human players (quartet). The motif is the shadow of an imaginary pianist. This imaginary pianist indicates the timing and the expected tempo to begin the next phrase with a pianist-like cue motion (up and down the arms with waving the body) driven by the system. Each player easily sees this visualization ahead of each music stand. The cue motion of the imaginary pianist can be controlled by two parameters. The first parameter is the amplitude that determines how vigorously the pianist moves. The second parameter is the duration of the motion generated by the pianist. Usually, the amplitude is set according to the strength of the next note. The duration can be set according to the playing beat duration (inverse tempo); in practice, we set it to 1.5 times the current beat duration by default. We annotate these parameters in the music score during the rehearsals, and during the performance, the system begins the motion when the playback position reaches the annotated position. Additionally, the imaginary pianist continuously undulates, driven by the starting of new notes by the player piano (please see the supplemental video); this roughly informs the players about the system's playback tempo.

Expression-Preserving Coordinator

In order to coordinate the timing, the system plays the following position $m(t)$ at time t:

$$m(t) = x_{C,n} + (L + t - \tau_n) v_{C,n}. \quad (6)$$

Variables $v_{C,n}$ and $x_{C,n}$ are the velocity and the temporal offset, respectively, obtained when the most recent note is played, either by the machine or humans, at time τ_n (*i.e.*, n notes have been played so far).

The system generates the velocity $v_{C,n}$ and offset $x_{C,n}$ by coordinating between where the humans are playing and where the machine "wants" to play.

First, in order to express where the human players are playing, we assume that humans play with a piece-wise constant

velocity $v_{P,n}$ between τ_n and τ_{n+1}. In other words, we express the score position played by human players as follows, where $x_{P,n}$ is the position played at time τ_n and $\varepsilon_{P,n} \in \mathbb{R}^2$ is an additive Gaussian noise:

$$x_{P,n} = x_{P,n-1} + \Delta T_{n,n-1} v_{P,n-1} + \varepsilon_{P,n,0} \quad (7)$$
$$v_{P,n} = v_{P,n-1} + \varepsilon_{P,n,1}. \quad (8)$$

Here, we let $\Delta T_{m,n} = \tau_m - \tau_n$. Additive noise $\varepsilon_{P,n}$ contains the change of tempo and the deviation of the timing from the piece-wise constant velocity assumption.

Second, in order to express how the machine "wants" to play, the model sets the tempo curve of the machine part about a "default" tempo trajectory. The system assumes the position $x_{M,n}$ and speed $v_{M,n}$ evolves as follows:

$$x_{M,n} = x_{M,n-1} + \Delta T_{n,n-1} v_{M,n-1} + \varepsilon_{M,n,0} \quad (9)$$
$$v_{M,n} = \beta_n v_{\bar{M},n} + (1-\beta) v_{M,n-1} + \varepsilon_{M,n,1}. \quad (10)$$

Here, $v_{\bar{M},n}$ is the default tempo at the position reported by the score follower at time τ_n, and $\varepsilon_{M,n}$ is an additive Gaussian noise. $\beta_n \in [0,1]$ is a parameter that governs how strongly the machine part "wants" to revert to the default tempo $v_{\bar{M},n}$. We call this the *machine-reverting parameter*. This model is similar to the coordination strategy in the preliminary design, except (1) the machine-reverting parameter is dependent on the score position and (2) the model describes the temporal dynamics of only the machine part.

Third, to coordinate the timing between the human playing and what the machine "wants" to play, the two timing models are coupled together. Specifically, the timing of the machine part is corrected by the human parts' predictions. This kind of model is inspired by the first-order, linear phase and period correction model used in the human perception of rhythmic coordination (sensorimotor coordination) [30]. We call the strength of correction the *coupling parameter* $\gamma_n \in [0,1]$. Given the coupling parameter, the playback position and the tempo at time τ_n is given as follows:

$$x_{C,n} = x_{M,n} + \gamma_n (x_{P,n} - x_{M,n}) \quad (11)$$
$$v_{C,n} = v_{M,n} + \gamma_n (v_{P,n} - v_{M,n}). \quad (12)$$

In this model, γ_n affects the extent of correction; at the extrema, the machine ignores the humans when $\gamma_n = 0$, and the machine tries to perfectly synchronize with the humans when $\gamma_n = 1$. The variance is a weighted combination of both the variance of the machine part $x_{M,n}$ and the humans $x_{P,n}$, allowing the uncertainties of both models to be mixed naturally. We bootstrap γ_n with a symbolic analysis of the music score, and allow γ_n to be overridden through manual annotation. Symbolic analysis provides the initial basis for coordination such as the clarity of rhythm [20, 12], and overriding incorporates the preferences of the human musicians, similar in spirit to music conducting systems [1]. Specifically, to bootstrap the coupling parameter γ_n from the musical score, we compute the density of the note onsets $\phi_{\mathbf{n}} \in \mathbb{R}^2$, where $\phi_{n,0}$ and $\phi_{n,1}$ contain the moving average of the note density of the machine part and the human part, respectively. We then assume that parts with more note onsets lead the ensemble, and set $\gamma_n = (\phi_{n,1} + \varepsilon)/(\phi_{n,1} + \phi_{n,0} + 2\varepsilon)$ for some small $\varepsilon > 0$.

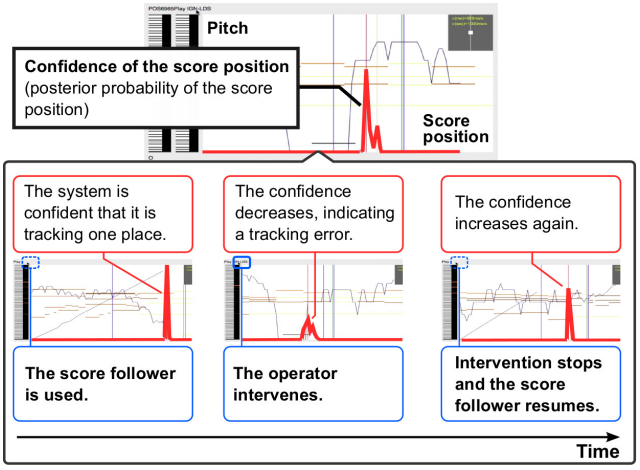

Figure 5. The operating view during the performance. The system displays the internal confidence and a backstage operator monitors it. If the internal confidence decreases, the operator can fix it (relevant UI elements are highlighted for clarity).

Finally, the timing reported by the score follower, *i.e.*, μ_n and σ_n, is incorporated as an observation from the coordinator:

$$[\mu_n, \mu_{n-1}, \cdots, \mu_{n-I_n}] \sim \mathcal{N}\left(\mathbf{W}_n [x_{P,n}, v_{P,n}], \mathrm{diag}([\sigma_n^2, \sigma_{n-1}^2, \cdots, \sigma_{n-I_n}^2])\right). \quad (13)$$

Here, I_n is the length of history considered, set such that all note events that have occurred one beat before τ_n are contained. $\mathbf{W}_n$ contains the linear prediction coefficients to predict μ_n from $x_{P,n}$ and $v_{P,n}$, given as follows:

$$\mathbf{W}_n^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \Delta T_{n,n} & \Delta T_{n,n-1} & \cdots & \Delta T_{n,n-I_n+1} \end{pmatrix}. \quad (14)$$

For a real-time inference during a live performance, the system updates the timing models when (1) receiving $(\tau_n, \mu_n, \sigma_n^2)$ from the score follower and (2) the machine part plays a new note. Since the model is linear-Gaussian, the coordinator may be updated highly efficiently as a Kalman filter [19]. When receiving $(\tau_n, \mu_n, \sigma_n^2)$ from the score follower, the system performs the predict and the update steps of the Kalman filter to update the state variables $\{x_{C,n}, v_{C,n}, x_{M,n}, v_{M,n}, x_{P,n}, v_{P,n}\}$. Furthermore, when the machine part plays a new note, the system replaces the state variables by the predicted values from the predict step of the Kalman filter.

Manual Intervention

A fully automatic or a fully manual music ensemble system is difficult to use in a concert. A fully automatic system is difficult because an automatic tracking algorithm may make critical tracking errors at few musical phrases that are inherently difficult to track, and ruin the concert. On the other hand, a fully manual system, say, through having a backstage human operator "tap" the beats of human musicians [9], is stressful for the operator [Ai, personal communication], more so in classical music compared to other genres of Western music such as jazz, since the tempo fluctuates more significantly and the duration of a piece tends to be longer.

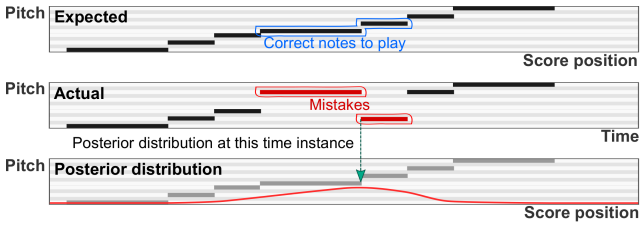

Figure 6. Illustration of a human-rooted error. When the human performer makes a mistake, the system becomes unsteady. Increased variance of the posterior distribution indicates this error.

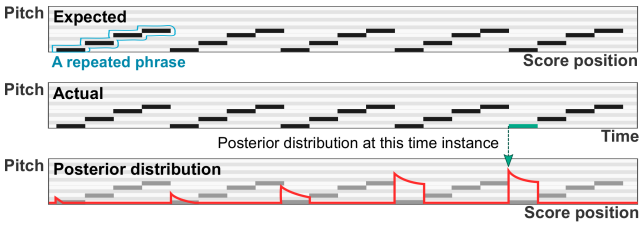

Figure 7. Illustration of a content-rooted error. A repeated phrase makes the system unstable. Multiple peaks in the posterior distribution indicates this error.

As a compromise, the system uses a semi-automatic ensemble system using human interventions. That is, to prevent a catastrophic failure of a concert due to the failure of automatic tracking, the system allows a backstage human operator to "intervene" in real-time. The system continuously displays the posterior distribution over the score position (Figure 5). By displaying the posterior distribution, key signs of failures may be visualized, allowing the operator to take preemptive measures. With this system, manual intervention compensates for the low reliability of a fully automatic system, and automatic tracking lightens the burden incurred to the operator of a fully manual system.

There are two main reasons for the system to fail, both of which are predictable. First, the system may fail because the human players take unexpected actions, such as playing wrong notes. Although the follower would use the temporal dynamics assumed by the HMM to keep track of the position, the system will eventually fail if wrong notes are played consecutively. We call this kind of failure a *human-rooted* failure. Human-rooted failure is predictable because the system loses confidence of the tracked position, seen as an increased variance of the posterior distribution (Figure 6). Thus, by monitoring the variance of the posterior distribution, the operator may take preventive measures.

Second, the system may fail because some segments of the music score is inherently difficult for the follower to track. We call this kind of failure a *content-rooted* failure. Content-rooted failure occurs, for example, in a short repetition; repetitions are difficult to track with a first-order HMM because it cannot "remember" which iteration of the repetition it was in. A content-rooted failure is predictable because the system becomes "confident" that it is tracking two or more positions simultaneously. That is, the posterior distribution exhibits many prominent peaks (Figure 7). Thus, by monitoring the number of prominent peaks in the posterior distribution, the operator may take preventive measures.

If a failure is about to occur or if anything sounds wrong to the ears of the operator, the operator may intervene in a few ways. First, a human-rooted error may be prevented by ignoring the input features, using only the temporal dynamics assumed by the HMM to track. Second, an operator may hold the position of the follower at a specified place, preventing the score follower from advancing. The position may be held by setting the likelihood of the current position to 1 and everything else to 0. Third, a content-rooted error may be prevented by ignoring the reported position of the score follower; this is useful if the score follower fails for a short time but recovers once a difficult-to-track segments are over. Fourth, the operator may directly adjust the tempo and the position of the coordinator.

RESULTS

We used *MuEns* in a real-world concert, with the support of the Tokyo University of the Arts. Since a quantitative evaluation is difficult due to the subjective and irreproducible nature of a live music performance, we discuss the system in the context of preparation of the concert and the actual concert. The repertoire and ensemble are the same as that employed in the preliminary study. For manual intervention, the system was operated by one of the authors. In this section, we denote the comments made by the violinist, violist, cellist and the bassist as "Vn," "Va," "Vc" and "Cb," respectively.

Rehearsal

We took three, one hour-long rehearsals during the three days into the concert. The rehearsals were videotaped and the conversations were transcribed.

During the rehearsal, the machine-reverting parameter and the coupling parameter were adjusted whenever the ensemble stopped playing and requested the behavior of the system to be changed. The machine-reverting parameter was adjusted based on how much "character" of the piano part needed to be retained. Thus, the ensemble initially commented on the musicality of the piano part, which was used to adjust the machine-reversion character; eventually, these comments were eliminated. The bootstrapped coupling parameter was sufficient in many cases. One of the exceptions included the first variation of the fourth movement, where the piano has the melody (low note density) and the ensemble accompanies (high note density). While the bootstrapping method would cause the ensemble to lead, the ensemble wanted the piano to lead instead. In this very passage, the absolute mean error of onset timing between the piano and the double bass (who leads the ensemble) decreased from 120 ms (using bootstrap parameter) to 40 ms (after adjusting the behavior).

As the timing issues were resolved, the ensemble went on to spend most of the rehearsal matching the nuance of the strings to the piano. As the ensemble tried to match the nuance, they requested some tempo and dynamics of the piano part to be adjusted. Note that this is unlike the preliminary experiment, where they were skeptical of "manipulations (Cb)."

For visual cue detection, the person giving the cue was agreed on during the rehearsal. For example, during the rehearsal,

Figure 8. Setup of the concert.

the ensemble discussed and agreed on the instrument that is responsible for recovering from a specific *fermata*.

For visual cue generation, the places to generate the cue have been decided by a pianist. The ensemble was initially startled by the gestures generated by the visualizer, which did not match the dynamics (volume) of the piano part:

> What really, really threw me off is that I think good pianist would never do the same kind of move whether piano [weak] or forte [strong] (Vn).

The ensemble initially thought that the visualizer is a distraction. However, after changing the extent of gesture to match the dynamics of the piano, the ensemble thought the visual feedback is "super (Cb)" and is useful for coordination. Indeed, the mean absolute difference of onset timing between the ensemble and the piano for the entry timings to selected portions that use visual feedback (parts of the fifth movement where everyone plays in unison) decreased from 180 ms (no visual feedback) to 70 ms (visual feedback).

For manual intervention, we found that the critical errors were all content-rooted and thus were predictable ahead of the performance. To plan the intervention strategy, the operator annotated parts of the piece where the system consistently failed and practiced the intervention operations. The operator practiced the intervention operations outside the actual rehearsal through simulation, *i.e.,* by feeding the system a recording of the rehearsal to the system and operating the system as to avoid failures. The practiced operation was executed during the rehearsals with the ensemble and the actual concert.

Concert

The concert was held on 19 May 2016 at the Tokyo University of the Arts. The setup of the concert was as shown in Figure 8, consisting of a piano quintet in a traditional formation, with the piano driven using a player piano. A projector was used to display the visual feedback to the musicians, showing the visualization inside a circle bounded by the four performers.

During the concert, the operator was standing by the stage, listening to the string players. Of about fifteen minutes of performance, there were about twenty seconds of manual interventions by the human operator. First, the operator ignored the score following output in about twelve bars of the piece, where a repeating sequence made the system highly prone to content-rooted errors. Second, the operator held the position that contains a long held note and is also prone to a content-rooted error. In addition, the operator adjusted the microphone input gain, so that the sound produced by the player piano won't adversely affect the tracking.

Reception of the Audience

The concert was marketed as an ensemble between Sviaslatov Richter resurrected using an artificial intelligence (AI) technology and the Scharoun Ensemble. Over five hundred people attended the concert.

There were about ten posts on microblogs, posted between the day of the concert and two days after, that mentioned about the quality of the synchronization. The posts suggest that the main audiences were (1) tech-savvy persons, interested in AI applied to music ensemble, and (2) classical music fans. Some posts mentioned that the synchronization between machine and humans gave "goosebumps," and that the timing was "fine." About half mention that the performance was "mundane" and "on the safe side." One post mentioned that the ensemble sounded that it was "trying" to keep up with the machine.

A newspaper review [17] mentioned that the audience "gasped" and "cheered" at the performance, and that a future work is keeping up with the "spontaneity" of the musicians.

Interview with the Musicians

We describe some comments made by the musicians, through informal conversations during the rehearsals.

The musicians found that the proposed system, after tuning, was "one thousand times better (Cb)" and that they have nothing to say about the behavior of the system.

When matching the nuance, the ensemble was confused when the timing nuance has changed unexpectedly. For example, the ensemble paused the rehearsal when the nuance of the piano part has changed, questioning, "why did it do this [change of nuance] (Vn)?" They thought that it is "weird (Cb)" that the nuance is "different – too different (Vn)."

Some factors pointed out by the ensemble are still unaddressed in the system. First, the ensemble commented that it is important to listen to a particular instrument:

> [Listening to the ensemble] might be sometimes dangerous because we [all but a particular instrument] may be doing strange things. It would be better to just react to [the particular instrument] in [a particular context] (Cb).

Indeed, the ensemble seemed to naturally listen for a particular player that keeps the time:

> "Good, have we have rhythm in the bass." This [kind of inference] is easy for humans (Cb).

Second, they found that it is a "stressful (Vn)" experience to play with a data with the "same touch [*i.e.*, sequence of note strengths] (Cb)" and no mistakes. To elaborate, they thought that they "were not allowed to make mistakes (Vn)" because the piano data was such. Third, the ensemble found that the system lacks "humor (Cb)." We believe that this means that the system responds with the same touch, and is incapable of responding to the nuance of the human performers.

Discussion

MuEns enabled the musicians to make music at a higher level than mere synchronization. That is, the issues initially raised during the rehearsal with the preliminary system was mostly regarding the synchronization and how the machine should respond. On the other hand, the issues addressed with the proposed system switched to higher musical issues, such as how the ensemble should play its parts in response to the piano part. This suggests that the proposed system provides sufficient coordination for the musicians to start making music at a higher level, that is, coordinating its nuances to the machine.

This kind of improvement is a combined effect of incorporating the multi-modal cue generation and recognition, the expression-aware timing model, and the scheme for manual intervention. First, the visual cue generation and detection enabled better coordination, allowing humans to coordinate the starting time of a phrase. The visualization was essential once the generated gesture was consistent with the generated piano sound, as suggested by both the musicians' comments and the decreased onset timing error. Second, the capability to adjust the machine-reverting and the coupling parameters enabled the musicians discuss and fix coordination strategies. These corrections improved the coordination of the ensemble, as suggested by the lowered onset timing error. Furthermore, the resolved timing issues allowed the musicians to discuss other issues at a higher musical level. Third, when the score follower failed, manual intervention allowed the musicians to continue on playing. This was important not only for the stage performance but also for the rehearsals, since if the piano part responded erroneously, the ensemble could not match its nuances with the piano.

The study shed light to further issues that need to be addressed in the future. The main theme of the issues is allowing top-level musicians to exhibit a high degree of artistic freedom.

First, we found that as the ensemble got more synchronized with the system, the music performance got stringent. This is perhaps the reason why the audiences thought the performance was on "the safe side." While this kind of reduced freedom occurs in inter-human ensemble as well [12], it seemed more prominent here, as the reviews suggest. We believe this is attributed to two causes. First, the performers, having encountered instances of bad synchronization in the preliminary design and during the rehearsals, got risk-averse and started playing safely. Second, the members of the ensemble have known each other for years, but know the system for only few hours. Thus, the members, when playing with each other, are capable of playing highly freely because they know each other's playing style well. On the other hand, when playing with the system, they still could not grasp the expected response of the system. Further investigation is necessary to see if human musicians, after using the system for a long time, could interact freely with a machine as they would with a well-acquainted musical partner.

Second, significant time during and between rehearsals was lost annotating, writing and re-loading the data. This happened because the proposed system annotates the coupling and the machine-reverting parameters to the quartet MIDI score data, using an external MIDI sequencer. Thus, the comments made by the ensemble needed to be adjusted in a sequencer, writing as a standard MIDI file, and re-loading the file on the ensemble system. In the future, tightly knit integration of the annotation and the ensemble system is essential.

Third, the musicians seem to want a system whose response is *consistently nuanced*, but not identically generated. Consistent response is important because the musicians adjust their playing style to the expected behavior of the system. For example, the ensemble stopped playing when the piano part responded highly inconsistently, wondering "why did it do this [change of nuance]." Thus, if the machine is inconsistent, it is impossible for humans to conform to the nuance of the machine. While seemingly contradictory, however, identical playback intimidates the musicians. For example, the members found the use of the system "stressful" because it produces an identical sequence of note strengths every time; this kind of consistency intimidated the musicians into thinking that they too cannot make any mistakes. Instead of an identical playback data, the musicians seem to want a variation in the data that is consistent with the musical idea behind the music performance. For example, the ensemble thought that the system felt "different" from a human, since it lacks the "humor" of a human musician. To elaborate, each player in a human ensemble infers the expressive intent of each other and responds accordingly – indeed, the ensemble adapted their style to match the Russian nuance of the piano data. On the other hand, the system is oblivious to the musical idea behind the human musicians' playing. Hence the system response fails to elaborate on the musical idea underlying the tracked timing.

CONCLUSION AND FUTURE WORK

This paper presented *MuEns*, a system for incorporating machines in a music ensemble performance. It uses audio-visual cues to track the human players and coordinates the player piano playback and the visual cue generation. Timing is coordinated such that it balances between how the system should play and how the system should synchronize to the humans. The system is useful in a live concert where tracking must never fail, thanks to the manual intervention mechanism, which allows a human operator to guide the system through hard-to-track passages. We have verified the system in a real-life concert scenario, and confirmed that the viability of system.

We address future works. First, intervention-friendly mathematical model is required, since the decoupling of the score follower and coordinator complicates the handling of manual intervention. Second, development of a system with tightly integrated annotation and playback is important. Since the time is limited in a professional rehearsal, the playback system should adapt itself quickly to newly annotated information. Third, responding with a variability of data is an important task, especially for relieving the stress incurred to human musicians. Finally, it is important to develop a system that understands and responds to the underlying intent of the human performer.

ACKNOWLEDGMENTS

This research is partially supported by the Center of Innovation Program from Japan and the Tokyo University of the Arts.

REFERENCES

1. Takashi Baba, Mitsuyo Hashida, and Haruhiro Katayose. 2010. "VirtualPhilharmony:" A Conducting System with Heuristics of Conducting an Orchestra. In *Proc. New Interfaces for Music Expression*, Vol. 2010. 263–270.
2. Julio José Carabias-Orti, Francisco J. Rodríguez-Serrano, Pedro Vera-Candeas, Nicolás Ruiz-Reyes, and Francisco J. Cañadas-Quesada. 2015. An Audio to Score Alignment Framework Using Spectral Factorization and Dynamic Time Warping. In *Proc. International Conference on Music Information Retrieval*. 742–748.
3. Ali T. Cemgil. 2009. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience* 2009 (2009).
4. Marcelo Cicconet, Mason Bretan, and Gil Weinberg. 2012. Visual Cues-based Anticipation for Percussionist-Robot Interaction. In *Proc. of the International Conference on Human-Robot Interaction*. 117–118.
5. Arshia Cont. 2008. ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music. In *Proc. International Computer Music Conference*. 33–40.
6. Arshia Cont. 2010. A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 6 (2010), 974–987.
7. Arshia Cont, Diemo Schwarz, Norbert Schnell, and Christopher Raphael. 2007. Evaluation of Real-Time Audio-to-Score Alignment. In *Proc. International Conference on Music Information Retrieval*. Vienna, Austria, 315–316.
8. Roger B Dannenberg. 1984. An On-line Algorithm for Real-time Accompaniment. In *Proc. International Computer Music Conference*. 193–198.
9. Roger B. Dannenberg. 2011. A Virtual Orchestra for Human-Computer Music Performance. In *Proc. International Computer Music Conference*. 185–188.
10. Roger B. Dannenberg and Christopher Raphael. 2006. Music Score Alignment and Computer Accompaniment. *Commun. ACM* 49, 8 (2006), 38–43.
11. Sebastian Ewert and Meinard Müller. 2012. Using Score-Informed Constraints for NMF-based Source Separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. 129–132.
12. Dorottya Fabian, Renee Timmers, and Emery Schubert (Eds.). 2014. *Expressiveness in Music Performance*. Oxford University Press.
13. Nicolas E. Gold, Octav-Emilian Sandu, Praneeth N. Palliyaguru, Roger B. Dannenberg, Zeyu Jin, Andrew Robertson, Adam Stark, and Rebecca Kleinberger. 2013. Human-Computer Music Performance: From Synchronized Accompaniment to Musical Partner. In *Proc. Sound and Music Computing Conference*. 277–283.
14. Lorin Grubb and Roger B Dannenberg. 1997. A Stochastic Method of Tracking a Vocal Performer. In *Proc. International Computer Music Conference*. 301–308.
15. Ning Hu, Roger B. Dannenberg, and George Tzanetakis. 2003. Polyphonic Audio Matching and Alignment for Music Retrieval. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*. 185–188.
16. Tatsuhiko Itohara, Kazuhiro Nakadai, Tetsuya Ogata, and Hiroshi G. Okuno. 2012. Improvement of Audio-Visual Score Following in Robot Ensemble with Human Guitarist. In *Proc. International Conference on Humanoid Robots*. 574–579.
17. Takayuki Iwasaki. 2016. AI Piano to Ningen ga Gassou – Kyoshou no Ensou, Saigen he Daiippo [AI Piano and Humans Plays in an Ensemble – One Step Forward for Recuperating Past Legend's Performance]. *The Nikkei [In Japanese]* (23 May 2016).
18. Cyril Joder, Slim Essid, and Gaël Richard. 2010. A Conditional Random Field Viewpoint of Symbolic Audio-to-Score Matching. In *Proc. ACM Multimedia*. 871–874.
19. Rudolph E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82, 1 (1960), 35–45.
20. Peter E. Keller. 2001. Attentional Resource Allocation in Musical Ensemble Performance. *Psychology of Music* 29, 1 (2001), 20–38.
21. Angelica Lim, Takeshi Mizumoto, Louis-Kenzo Cahier, Takuma Otsuka, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. 2010. Robot Musical Accompaniment: Integrating Audio and Visual Cues for Real-time Synchronization with a Human Flutist. In *Proc. Intelligent Robots and Systems*. 1964–1969.
22. Akira Maezawa, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Hiroshi G. Okuno. 2015. Unified Inter- and Intra-Recording Duration Model for Multiple Music Audio Alignment. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*. 1–5.
23. Nicola Montecchio and Arshia Cont. 2011. Accelerating the Mixing Phase in Studio Recording Productions By Automatic Audio Alignment. In *Proc. International Conference on Music Information Retrieval*. 627–632.
24. Takuma Otsuka, Kazuhiro Nakadai, Tetsuya Ogata, and Hiroshi G. Okuno. 2011. Incremental Bayesian Audio-to-Score Alignment with Flexible Harmonic Structure Models. In *Proc. International Conference on Music Information Retrieval*. 525–530.
25. François Pachet. 2002. The Continuator: Musical Interaction with Style. In *Proc. International Computer Music Conference*. 211–218.
26. Miller Puckette and Cort Lippe. 1992. Score Following in Practice. In *Proc. International Computer Music Conference*. 182–182.

27. Christopher Raphael. 2001. A Bayesian Network for Real-Time Musical Accompaniment. In *Proc. Advances in Neural Information Processing Systems*. 1433–1439.

28. Christopher Raphael. 2004. A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores. In *Proc. International Conference on Music Information Retrieval*. 387–394.

29. Christopher Raphael. 2010. Music Plus One and Machine Learning. In *Proc. International Conference on Machine Learning*. 21–28.

30. Bruno H. Repp. 2005. Sensorimotor Synchronization: a Review of the Tapping Literature. *Psychonomic Bulletin & Review* 12, 6 (2005), 969–992.

31. Bogdan Vera, Elaine Chew, and Patrick G. T. Healey. 2013. A Study of Ensemble Synchronisation Under Restricted Line of Sight. In *Proc. International Conference on Music Information Retrieval*. 293–298.

32. Guangyu Xia and Roger B. Dannenberg. 2015. Duet Interaction: Learning Musicianship for Automatic Accompaniment. In *Proc. New Interfaces for Music Expression*. 259–264.

33. Guangyu Xia, Yun Wang, Roger B. Dannenberg, and Geoffrey Gordon. 2015. Spectral Learning for Expressive Interactive Ensemble Music Performance. In *Proc. International Conference on Music Information Retrieval*. 816–822.

34. Xiao Xiao and Hiroshi Ishii. 2011. Duet for solo piano: MirrorFugue for Single User Playing with Recorded Performances. In *CHI Extended Abstracts*. 1285–1290.

35. Ryuichi Yamamoto, Shinji Sako, and Tadashi Kitamura. 2013. Robust On-line Algorithm for Real-time Audio-to-Score Alignment based on a Delayed Decision and Anticipation Framework. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. 191–195.