Noname manuscript No.
(will be inserted by the editor)

Interactive Improvisation with a Robotic Marimba Player

Guy Hoffman · Gil Weinberg

Received: date / Accepted: date

Abstract *Shimon* is a interactive robotic marimba player, developed as part of our ongoing research in Robotic Musicianship. The robot listens to a human musician and continuously adapts its improvisation and choreography, while playing simultaneously with the human. We discuss the robot's mechanism and motion-control, which uses physics simulation and animation principles to achieve both expressivity and safety. We then present an interactive improvisation system based on the notion of *physical gestures* for both musical and visual expression. The system also uses anticipatory action to enable real-time improvised synchronization with the human player.

We describe a study evaluating the effect of embodiment on one of our improvisation modules: antiphony, a call-and-response musical synchronization task. We conducted a 3x2 within-subject study manipulating the level of embodiment, and the accuracy of the robot's response. Our findings indicate that synchronization is aided by visual contact when uncertainty is high, but that pianists can resort to internal rhythmic coordination in more predictable settings. We find that visual coordination is more effective for synchronization in slow sequences; and that occluded physical presence may be less effective than audio-only note generation.

Finally, we test the effects of visual contact and embodiment on audience appreciation. We find that visual contact in joint Jazz improvisation makes for a performance in which audiences rate the robot as playing better, more like a human, as more responsive, and as more inspired by the human. They also rate the duo as better synchronized, more coherent, communicating, and coordinated; and the human as more inspired and more responsive.

Keywords Human-robot interaction · Robotic musicianship · Musical robots · Embodied cognition · Gestures · Anticipation · Joint action · Synchronization · User studies

G. Hoffman and G. Weinberg
Georgia Institute of Technology
Center for Music Technology
840 McMillan St, Atlanta, GA, 30332
E-mail: ghoffman@gmail.com , gilw@gatech.edu

1 Introduction

This paper describes Shimon, an interactive robotic marimba player. Shimon improvises in real-time while listening to, and building upon, a human pianist's performance. We have built Shimon as a new research platform for Robotic Musicianship (Weinberg and Driscoll, 2006b). As part of this research, we use the robot to evaluate some core claims of Robotic Musicianship. In particular, we test the effects of embodiment, visual contact, and acoustic sound on musical synchronization and audience appreciation.

We also introduce a novel robotic improvisation system. Our system uses a physical gesture framework, based on the belief that musicianship is not merely a sequence of notes, but a choreography of movements. While part of the function of these movements is to produce musical sounds, they also perform visually and communicatively with other band members and with the audience.

A physical motion based improvisation framework also extends traditional notions of computer-generated improvisation, which have usually put abstract note generation in the foreground. Our approach suggests a novel way to achieve real-time joint improvisation between a human and a machine, stemming from physical action, based on principles of embodiment and non-abstract cognition.

We also extend prior research on the use of an anticipatory action model for human-robot fluency, and integrate a similar anticipatory approach in our improvisation system, with the aim of promoting real-time musical coordination.

Our system was implemented on a full-length human-robot Jazz duet, displaying highly coordinated melodic and rhythmic human-robot joint improvisation. We have

performed with the system in front of live public audiences.

An abridged version of the motor control and improvisation system was outlined in a previous paper (Hoffman and Weinberg, 2010). This paper extends the technical description of the robotic system, and describes two new human subject studies evaluating the thus far unexplored effects of embodiment and visual contact on Robotic Musicianship.

1.1 Robotic Musicianship

We define Robotic Musicianship to extend both the tradition of computer-supported interactive music systems, and that of music-playing robotics (Weinberg and Driscoll, 2006b):

Most computer-supported interactive music systems are hampered by not providing players and audiences with physical cues that are essential for creating expressive musical interactions. For example, in humans, motion size often corresponds to loudness, and gesture location to pitch. These cues provide visual feedback and help players anticipate and coordinate their playing. They also create a more engaging experience for the audience by providing a visual connection to the sound. Most computer-supported interactive music systems are also limited by the electronic reproduction and amplification of sound through speakers, which cannot fully capture the richness of acoustic sound (Rowe, 2001).

On the other hand, much research in musical robotics focuses mostly on sound production alone, and rarely addresses perceptual and interactive aspects of musicianship, such as listening, analysis, improvisation, or interaction. Most such devices can be classified in one of two ways: the first category includes robotic musical instruments, which are mechanical constructions that can be played by live musicians or triggered by pre-recorded sequences (Singer et al, 2003; Dannenberg et al, 2005). More recently, Degallier et al (2006) demonstrated a nonlinear dynamical system for generating drumming trajectories in real time. Their system allows a robotic drummer to automatically switch and synchronize between discrete and rhythmic movements, but also does not address human musical input as part of the interaction. The second group includes anthropomorphic musical robots that attempt to imitate the action of human musicians (Solis et al, 2009; Toyota, 2010). Some systems use the human's performance as a user-interface to the robot's performance (Petersen et al, 2010); and only a few attempts have been made to develop perceptual, interactive robots that are controlled by autonomous methods (Baginsky, 2004).

One such system by Ye et al (2010) supports human-robot turn taking interaction using a multi modal approach. Their marimba playing robot detects volume decrease from human musical input, which triggers a vision system to detect a human head nod to approve and finalize the turn taking. The project, however, does not address human-robot temporal synchronization or joint improvisation. Lim et al (2010) developed a vision-based ensemble synchronization system that builds on existing score following techniques by analyzing periodic body movement to detect beat and tempo. The robot can dynamically synchronize its pre recorded score to the human generated beat and tempo.

In our previous work, we have developed a perceptual and improvisatory robotic musician in the form of *Haile*, a robotic drummer (Weinberg and Driscoll, 2006a). However, Haile's instrumental range was percussive and not melodic, and its motion range was limited to a small space relative to the robot's body. We have addressed these limitations with *Shimon*, a robot that plays a melodic instrument—a marimba—and does so by covering a larger range of movement (Weinberg and Driscoll, 2007).

2 Robotic Platform

Several considerations informed the physical design of *Shimon*: we wanted large movements for visibility, as well as fast movements for high note density. In addition our goal was to allow for a wide range of sequential and simultaneous note combinations. The resulting design was a combination of fast, long-range, linear actuators, and two sets of rapid parallel solenoids, split over both registers of the instrument.

Figure 1 shows two views of the robot. It is comprised of four arms, each actuated by a voice-coil linear actuator at its base, and running along a shared rail, in parallel to the marimba's long side. The robot's trajectory covers the marimba's full 4 octaves. Figure 2 shows a top-down diagram depicting the relationship between the linear actuator, arms, and instrument. The linear actuators are based on a commercial product by IAI and are controlled by a SCON trajectory controller. They can reach an acceleration of $3g$, and—at top speed—move at approximately one octave per 0.25 seconds.

The arms are custom-made aluminum shells housing two rotational solenoids each, drawn in Figure 3. The solenoids control mallets, chosen with an appropriate softness to fit the area of the marimba that they are most likely to hit. Each arm contains one mallet for the bottom-row ("white") keys, and one for the top-row

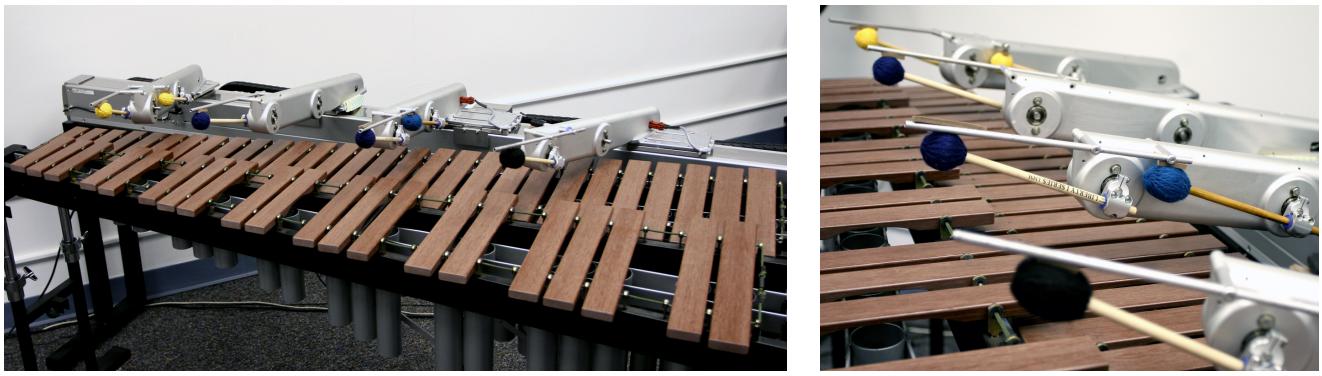

Fig. 1: Overall view (left) and detail view (left) of the robotic marimba player *Shimon*. Four arms share a voice-coil actuated rail. Two rotational solenoids per arm activated mallets of varying firmness.

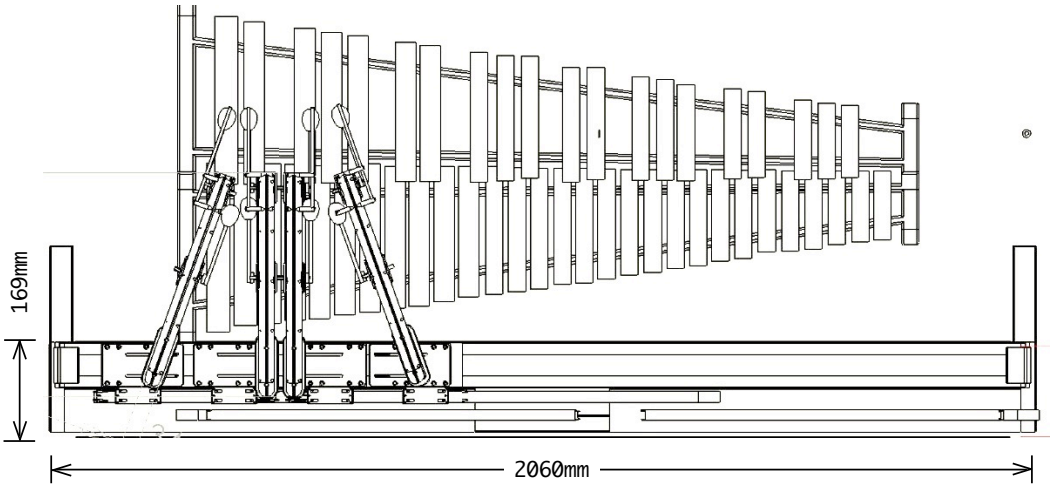

Fig. 2: Overall diagram of the relationship between the linear actuator, arms, and instrument of the robotic marimba player *Shimon*.

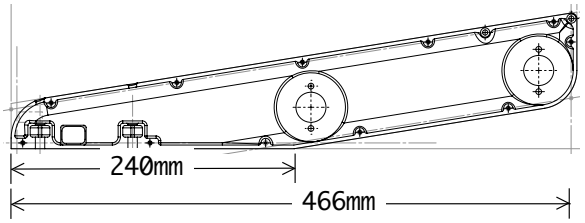

Fig. 3: Diagram showing a single mallet control arm, including the locations of the solenoid actuators (crosshairs).

("black") keys. *Shimon* was designed in collaboration with Roberto Aimi of *Alium Labs*.

3 Motor Control

A standard approach for musical robots is to handle a stream of MIDI notes and translate them into actuator movements that produce those notes. In *Shimon*'s case, this would mean a note being converted into a slider movement and a subsequent mallet strike. Two drawbacks of this method are (a) an inevitable delay between activation and note production, hampering truly synchronous joint musicianship, and (b) not allowing for expressive control of gesture-choreography, including tonal and silent gestures.

We have therefore separated the control for the mallets and the sliders to enable more artistic freedom in the generation of musical and choreographic gestures, without compromising immediacy and safety.

Figure 4 shows the overall communication and control structure of the robot. The computer ("PC") separately controls the mallets through the Mallet Motor

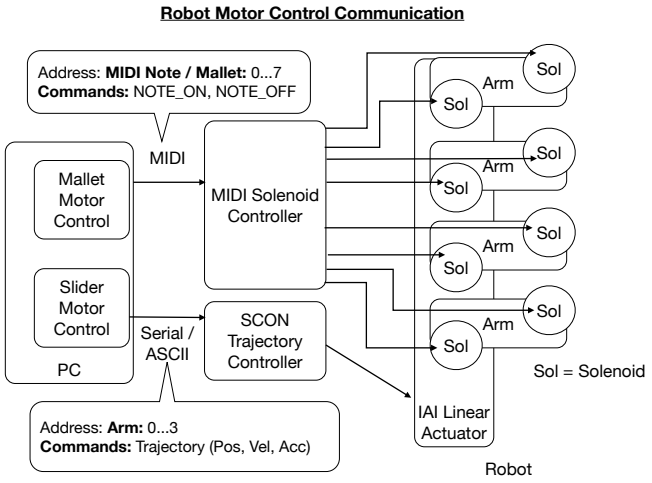

Fig. 4: Overall motor control communication diagram of the robot *Shimon.*

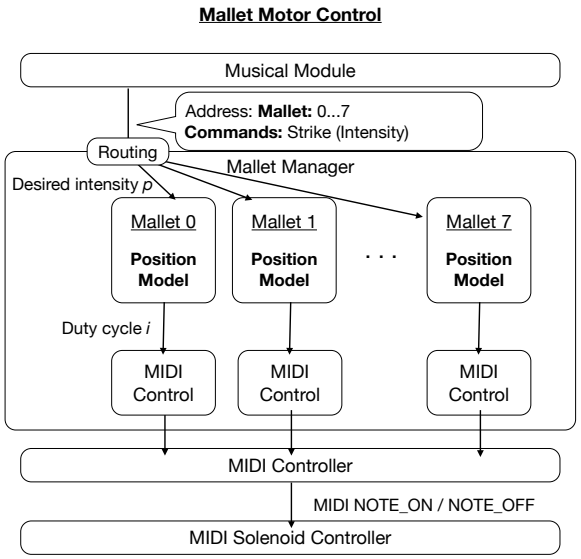

Fig. 5: Mallet Motor Control module diagram of the robot *Shimon.*

Control module (MMC), and the Slider Motor Control module (SMC). Both modules are further described in the sections below.

The MMC generates MIDI `NOTE_ON` and `NOTE_OFF` commands addressed to each of the 8 mallet rotational solenoids. These commands are demultiplexed by the MIDI Solenoid controller to actuator currents. The SMC uses IAI's proprietary SCON/ASCII serial protocol to specify slider positions and motion trajectories for each of the four linear actuators (sliders).

The remainder of this section describes the structure of the MMC and SMC control systems, which were designed with safe, yet artistic expressivity in mind.

3.1 Mallet Motor Control

The mallets are struck using rotational solenoids responding to on/off control through a MIDI Solenoid Controller. Eight arbitrary MIDI notes are mapped to the eight mallets, and the MIDI `NOTE_ON` and `NOTE_OFF` messages are used to activate and deactivate the solenoid.

Given this binary discrete electro-mechanical setup, we still want to be able to achieve a large dynamic range of striking intensities (i.e. soft and loud notes). We also want to be able to strike repeatedly at a high note rate.

This is achieved using the Mallet Motor Control module (Figure 5). Its goal is to translate a dynamic range of intensities for each mallet strike into a timing of the MIDI `NOTE_ON` and `NOTE_OFF` commands sent to the MIDI Solenoid controller. Note that this figure corresponds to the boxes labeled "Mallet Motor Control" and "MIDI Solenoid Controller" in Figure 4. The core of the motor control module is a system that translates the desired intensity p for each mallet into the MIDI duty cycle i.

As we can only control the solenoids in an on/off fashion, the striking intensity is a function of two parameters: (a) the velocity gained from the distance traveled; and (b) the length of time the mallet is held on the marimba key.

To calculate the appropriate duty cycle, we therefore need to maintain a model of the mallet position for each striker, and determine the correct duty cycle per position and mallet. In order to do so, we have empirically sampled sound intensity profiles for different solenoid activation lengths, and used those to build a base model for each striker (Figure 6). This model includes four parameters:

- $d_{\downarrow}$ — the mean travel time from the rest position to contact with the key,
- $d_{\uparrow}$ — the mean travel time from the down position back to the rest position,
- $d_{\rightarrow}$ — the hold duration that results in the highest intensity note for that particular mallet, and
- i_m — the duty cycle that results in the highest intensity note for that mallet, when it starts from the resting position.

Using this model, each of the eight mallet control modules translates a combination of desired strike intensity and time of impact into a solenoid duty cycle and trigger time. Intuitively—the lower a mallet is at request time, the shorter the duty cycle needs to be to reach impact, and to prevent muting of the key through a prolonged holding time.

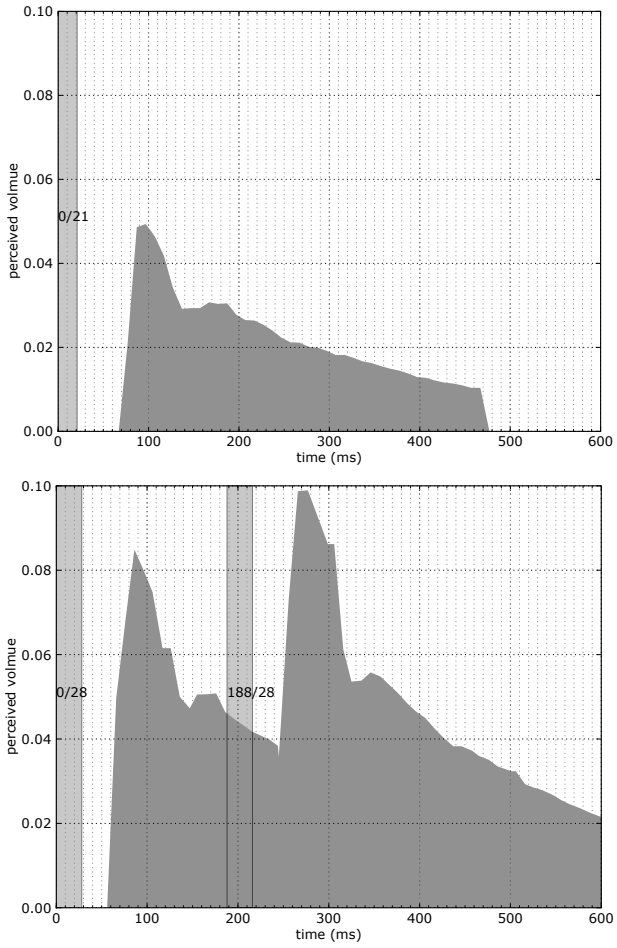

Fig. 6: Empirical strike/sound measurements used to build mallet models. We show one example each for single strike measurement to estimate $d_\downarrow$, $d_\rightarrow$, and i_m (top), and dual strike measurements used to estimate $d_\uparrow$ (bottom).

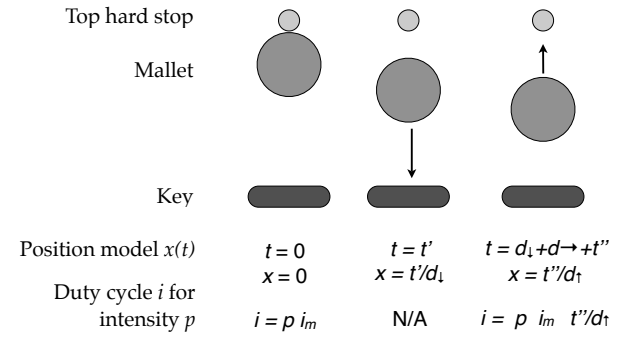

Fig. 7: Duty-cycle computation based on mallet position model

An estimated position x is thus dynamically maintained based on the triggered solenoid commands, and the empirical mallet model (Figure 7). During up-travel, $x(t)$, with t being the time since the last mallet activation start, is estimated as

$$x(t) = \frac{t - d_\downarrow - d_\rightarrow}{d_\uparrow} \quad (1)$$

As a result, the updated duty cycle i of mallet m as a function of the desired intensity p, is then

$$i = p \times i_m \times x(t) \quad (2)$$

The MIDI Controller then sends the appropriate `NOTE_ON` and `NOTE_OFF` commands on a separate program thread.

In the above equation, we approximate the mallet position as a linear function of travel time. Obviously, a more realistic model would be to take into account the acceleration of the mallet from the resting position to the key impact. Also, bounce-back should be accounted for, for short hold times. We leave these improvements for future work.

The described system results a high level of musical expressivity, since it (a) maintains a finely adjustable dynamic striking range from soft to loud key strokes, and (b) allows for high-frequency repetitions for the same mallet, during which the mallet does not travel all the way up to the resting position before being re-triggered.

3.2 Slider Motor Control

The horizontally moving sliders are four linear carriages sharing a rail and actuated through voice coil actuators under acceleration- and velocity-limited trapezoid control. This is done by the component labelled "SCON Trajectory Controller" in the diagrams herein.

There are two issues with this control approach. (a) a mechanical ("robotic"—so to speak) movement quality associated with the standard fire-and-forget motion control approach, and (b) collision-avoidance, since all four arms share one rail.

3.2.1 Animation Approach

To tackle these issues, we chose to take an *animation* approach to the gesture control. Based on our experience with previous robots, e.g. (Hoffman et al, 2008; Hoffman and Breazeal, 2004), we use a high-frequency controller running on a separate program thread, and updating the absolute position of each slider at a given frame rate (in most of our performances we use 40Hz). This controller is fed position data for all four arms

at a lower frequency, based on higher-level movement considerations.

This approach has three main advantages: (a) for each of the robotic arms, we are able to generate a more expressive spatio-temporal trajectory than just a trapezoid, as well as add animation principles such as ease-in, ease-out, anticipation, and follow-through (Lasseter, 1987); (b) since the position of the sliders is continuously controlled, collisions can be avoided at the position request level; and (c) movements are smooth at a fixed frequency, freeing higher-level to vary in update frequency due to musical or behavior control considerations, or processing bottlenecks.

This animation system is indicated at the bottom of Figure 9 above the communication layer to the SCON Trajectory Controller.

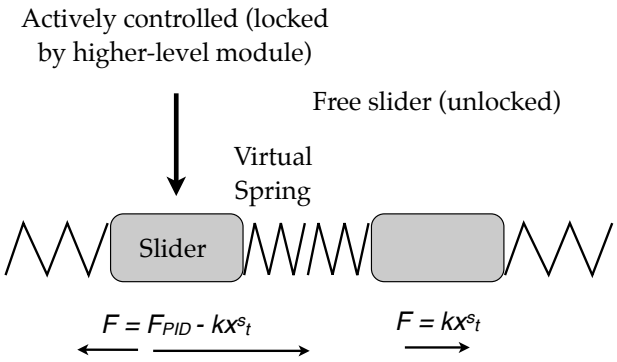

Fig. 8: Interaction between PID control and simulated spring model

3.2.2 Slider Manager: PID and simulated springs

The intermediate layer handling the slider position requests and generating the positions for each of the four sliders, while maintaining collision safety, is called the *Slider Manager*. It allows higher-level modules to "lock" a slider, and thus control it for the duration of the locking period.

The Slider Manager then uses a combination of PID control for each slider, with a simulated spring system between sliders, to update the position of all four sliders during each update cycle (Figure 8). For each position request x_t^r of a locked slider at position x_t at time t, we first calculate the required PID force using the discrete PID formula:

$$F_{PID} = K_p e_t + K_i \sum_{0}^{k} e_{t-i} d\tau + K_d \frac{e_t - e_{t-1}}{d\tau} \tag{3}$$

where $d\tau$ is the inverse sampling frequency, and

$$e_t = x_t^r - x_t$$

For sliders that are not locked, the PID force is 0.

In addition to the PID force, the Slider Manager models "virtual springs" on each side of each slider, which help prevent collisions and move unlocked sliders out of the way in a naturally-seeming fashion. Based on the current position of the carriages, the heuristically determined spring constant k, the length of the virtual springs, and thus their current simulated compression x_t^s at time t, we add the spring component kx_t^s to the force. The force update for each carriage is then

$$F_{PID} - kx_t^s \tag{4}$$

where the sign of kx_t^s for each spring is determined by the side of the simulated spring.

The result of this control approach is a system that is both safe—carriages will never collide and push each other out of the way—and expressive.

Figure 9 shows an overview of the Slider Motor Control module discussed in this section. This diagram corresponds to the boxes labeled "Slider Motor Control" and "SCON Trajectory Controller" in Figure 4. In sum, higher-level musical modules can "lock" and "unlock" each slider, and can request target positions for each locked slider. These target positions are translated through the PID controller for each slider into virtual forces, which are then combined for safety in the simulated spring resolver. The combined forces are used to update the target position of each arm. As the temporal resolution of this process is unpredictable and variable in frequency (in our applications, usually between 10-30Hz, depending on the musical application), these positions are transferred to the animation system, which runs on a separate thread at a fixed frequency (we normally use 40Hz) and updates the final motor position for each actuator using interpolation with velocity and acceleration limiting. The output positions from the animation controller are transmitted through the SCON ASCII serial protocol to the SCON Trajectory Controller.

While it can normally be assumed that higher-level modules will not cross over carriages, and be generally coordinated, adding this middle-layer control system allows more freedom of expressivity on a higher-level (for example inserting pre-scripted animations, or parametric gestures that control only a subset of the sliders), while still keeping the system safe and expressive at all times.

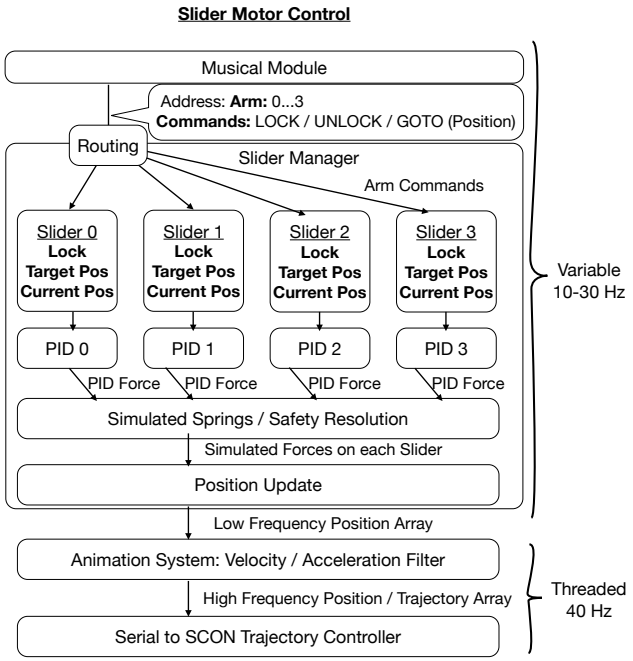

Fig. 9: Slider Motor Control module diagram of the robot *Shimon.*

4 Gestures and Anticipation

A main contribution of this paper is modeling interactive musical improvisation as *gestures* instead of as a sequence of notes. Using gestures as the building blocks of musical expression is particularly appropriate for robotic musicianship, as it puts the emphasis on physical movement and not on symbolic note data. Gestures have been the focus of much research in human musicianship, often distinguishing between tone-generating and non-tone-generating gestures (Cadoz and Wanderley, 2000). A physical-gestural approach is also in line with our embodied view of human-robot interaction (Hoffman and Breazeal, 2006), and similar perception-action based models of cognition.

Our definition of gesture deviates from the common use of the word in human musicianship. In this paper, a "gesture" is a physical behavior of the robot which may or may not activate the instrument, and can encompass a number of concurrent and sequential motor activities. A gesture could be a cyclical motion of one arm, a simple "go-to and play" gesture for a single note, or a rhythmic striking of a combination of mallets. Since gestures are not determined by notes, but by the robot's physical structure, *Shimon*'s gestures separately control the timing of the mallet strikers and the movement of the sliders, through the two motor control modules described in the previous sections. Improvisation and musical expression are thus in many ways the *result* of these physical gestures, rather than serving in their traditional role as amodal models that drive physical movement. The following section describes some gestures developed as part of this work, illustrating this notion.

4.1 Anticipatory Action

To allow for real-time synchronous non-scripted playing with a human, we also take an anticipatory approach, dividing each gesture into *preparation* and *follow-through.* This principle is based on a long tradition of performance, such as ensemble acting (Meisner and Longwell, 1987), and has been explored in our recent work, both in the context of human-robot teamwork (Hoffman and Breazeal, 2008), and for human-robot joint theater performance (Hoffman et al, 2008).

By separating the—potentially lengthy—preparatory movement (in our case: the horizontal movement) from the almost instant follow-through (in our case: the mallet action), we can achieve a high level of synchronization and beat keeping without relying on a complete-musical-bar delay of the system. Specifically, since for all mallets the travel time is below 90ms, the system operates at a < 100ms delay even for complex musical gestures involving horizontal travel often longer than 1 sec.

5 Improvisation

Implementing this gesture- and anticipation-based approach, we have developed a Jazz improvisation system, which we employed in a human-robot joint performance.

5.1 Background: Joint Jazz Improvisation

As the control system of *Shimon* described here is aimed at joint Jazz improvisation, this section provides a brief background on this style of musicianship.

The type of classic ("standard") Jazz addressed here is structured as a pre-set agreed-upon progression of chords with an associated melody superimposed on the chord progression. Most Jazz standard pieces (or simple "standards") have a relatively small number of chords in their progression. Each chord corresponds to a subset of a small number of possible scales, with possible adaptations of the subset, for tension and variation. Another way to think about a Jazz chord, is as a series of intervallic relationships to the root of the chord.

Jazz improvisation is usually structured around segments, often with the melody segment opening and closing the session. The melody segment generally has one or more instruments play the main melody according to the standard, with other instruments—referred to as the rhythm or accompaniment sections— accompanying the lead player with the harmonies specified by the chord progression, in synchrony to the melody.

A joint improvisation segment mostly still adheres to the chord progression, but can be flexible in terms of repetition, ordering, and length of each chord section. Within these sections, players are coordinated as to the current chord and tempo, but have relatively large freedom as to what melody and rhythm they use within the segment. That said, players are expected to use each others' playing as inspiration, and create a "back-and-forth" of sort, coordinating and mutually influencing each others' performance.

The improvisation system described in this paper attempts to model some of these aspects of standard Jazz joint improvisation, in particular the coordination of chords, tempo, and beats, the mutual "inspiration" between human and robot, the relative freedom of melody based on a certain chord, the separation into performance segments, and the standardized progression and selection of chords. Naturally, Jazz improvisation encompasses much more than these elements, and has a long tradition of particular musical structures, formats and sub-genres, which *Shimon* does not achieve. We believe, though, that the system discussed here provides a number of novel achievements in the realm of real-time joint Jazz improvisation between a human and a robotic player.

5.2 Human-Robot Joint Improvisation

In our system, a performance is made out of *interaction modules* each of which is an independently controlled segment or phase in the performance. It is continuously updated until the part's end condition is met. This is usually a perceptual condition, such as a chord played, or a certain tempo achieved, but can also be a pre-set amount of bars to play.

Figure 10 shows the general structure of an interaction module. It contains a number of gestures which are either triggered directly, or registered to play based on the current beat, as managed by the beat keeper. To recap: a gesture is a behavior module controlling zero or more sliders and mallets.

Gestures are selected and affected either by the beat (through the Beat Keeper module, described below), or by information coming in from percepts, which analyze

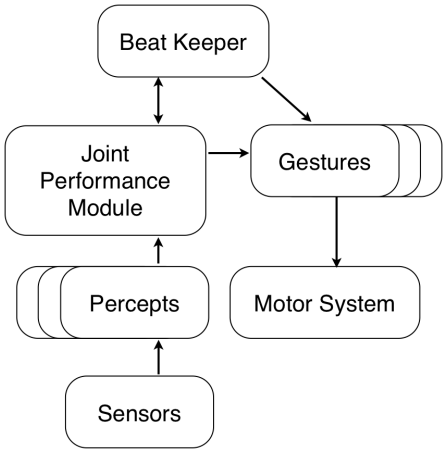

Fig. 10: Schematic interaction module for each phase of the performance

input from the robot's sensory system. These percepts can include, for example, a certain note density, or the triggering of a particular phrase or rhythm.

In this diagram, the box labeled "Gestures" corresponds to the "Musical Module" in the Motor Control diagrams in Figures 9 and 5, and the "Motor System" corresponds both the Slider Motor Control module, and the Mallet Motor Control module discussed above.

5.3 Improvisation Infrastructure

A number of infrastructure components are used by all improvisation modules.

5.3.1 MIDI Listener

While there are a number of sensory modules possible, we are currently using a MIDI sensory input, responding to the notes from a MIDI-enabled electric piano. On top of this sensor, we developed several perceptual modules described later in this section.

5.3.2 Beat Keeper

Common to all parts, and continuously running is the *Beat Keeper* module, which serves as an adjustable metronome that can be dynamically set and reset during play by the interaction module, and calls registered callback functions in the the gestures making up the performance.

To provide a simple example: a "one-three-random-note" gesture could register to get called on every "one" and "three" of a bar. In between calls it would "prepare" into a certain random position, and then would

activate the appropriate striker on the callback registered beats.

5.3.3 Chord Representation

We use three kinds of representations for Jazz chords in our system. The simple representation is that of a fixed set of notes in the robot's playing range. The second representation is octave-agnostic and includes a set of notes and all their octave harmonics. Finally, we also represent chords as a base note with a set of set-octave or octave agnostic harmonics.

5.4 Module I: Call-and Response

The first interaction module is the phrase-call and chord-response module. Call-and-response (sometimes called "antiphony") is a common musical interaction in joint musicianship. In this interaction, two musicians play two distinct musical phrases, where the second phrase is a commentary on the first phrase.

In order to enable the robot to play an appropriate "response" to the human's "call", one of the basic requirements is that the response is beat-matched and synchronized to the human's playing, i.e. that it starts on time, without delay, and that it plays in the correct tempo after it starts.

In this module, the system responds to a musical phrase with a pre-set chord sequence, arranged in a particular rhythmic pattern. The challenge here is not to select the right notes, but to be able to respond in time and play on a seamlessly synchronized beat and onset to that of the human player, who can vary the tempo at will.

This module makes use of the anticipatory structure of gestures. During the sequence detection phase, the robot prepares the chord gesture. When the phrase is detected, the robot can strike the response almost instantly, resulting in a highly meshed musical interaction.

This module includes two kinds of gestures:

Simple chord gestures —
: select an arm configuration based on a given chord during the preparation stage, and strike the prepared chord in the follow-through stage. If the chord is a set chord, the configuration is set. If it is a flexible chord, the gesture will pick a different arbitrary configuration satisfying the chord each time.

Rhythmic chord gestures —
: are similar to the simple chord gestures in preparation, but during follow-through will strike the mallets in a non-uniform pattern. This can be an arpeggiated sequence, or any other rhythmic structure.

The robot adapts to the call phrase using a simultaneous *sequence spotter* and *beat estimator* percept. Using an on-beat representation of the sequences that are to be detected, we use a Levenshtein distance metric (Levenshtein, 1966) with an allowed distance $d = 1$ to consider a phrase detected[1].

At that stage, the beat estimator will estimate both the played beat based on the duration of the sequence, and the beat synchronization based on the time of the last note played. The beat (in BPM) is calculated as follows:

$$bpm = \frac{60}{d_c / l_p} \tag{5}$$

where d_c is the duration of the call phrase in seconds, l_p is the length of the match phrase in beats

This value, as well as the synchronization estimate are transmitted to the beat keeper, which—through the above-mentioned beat callback mechanism—will cause execution of a sequence of simple and rhythmic chords. The result is an on-sync, beat-matched call-and-response pattern.

5.5 Module II: Opportunistic Overlay Improvisation

A second type of interaction module is the *opportunistic overlay improvisation*. This interaction is centered around the choreographic aspect of movement with the notes appearing as a "side-effect" of the performance. The intention of this module is to play a relatively sparse improvisation that is beat-matched, synchronized, and chord-adaptive to the human's playing.

The central gesture in this module is a rhythmic movement gesture that takes its synchronization from the currently active beat in the beat keeper module. An example of such a gesture would be a fixed periodic movement of each arm between two pre-set points. In the performance described below, we used an "opening and closing" gesture, in which the lower two arms always move in the opposite direction as the upper two arms, meeting in the middle of the instrument and turning around at the instrument's edges. As mentioned above, this movement is matched to the currently estimated human beat and tempo.

This beat is updated through a beat detection percept tracking the beat of the bass line in the human playing, using a simple bass-note temporal interval difference, modeled as either a full, a half, or a quarter bar based on the previous beat. In parallel, registering

[1] Naturally, we do not allow the last note in the phrase to be deleted for the purposes of comparison, as this would invalidate the synchronization

the human bass notes in such a way provides the down beat for each bar.

In parallel, a chord classification percept is running, classifying the currently played chord by the human player, by finding a best fit from the chords that are part of the current piece. Since chord classification is not central to this project, in the work presented here, we use only the bass notes from the beat detection percept to select among the chords that are part of the played piece. Furthering this work, we are considering a sliding window Bayesian Inference method for more flexible adaptation.

Without interrupting the periodic choreographic gesture, this interaction module attempts to opportunistically play notes that belong to the currently detected chord, based on a pre-programmed rhythmic pattern. For a quantized bar with m quantization bins, a rhythmic pattern is defined as the vector $(s_1, s_2, \cdots, s_m)$, where s_i is an indication of intensity $0 \leq s_i \leq 1$.

For each quantized time $0 \leq t \leq m-1$, if $s_t > 0$, the module checks if the horizontal position of one or more of the mallets corresponds to a key which is included in the currently detected chord. If a mallet matches this requirement, it strikes using the intensity s_t.

Since both the choreographic gesture and the rhythmic strike pattern are coordinated through a shared beat keeper, the result is a dynamically changing confluence of two rhythms and one chord structure, resulting in a novel improvisational gesture which is highly choreographic, and due to its complex "musical interference" structure can probably only be calculated and performed by a machine, but yet is still tightly synchronized to the human's playing, both in beat and harmony. This module exemplifies our argument for physical-motion and gesture based improvisation as an appropriate methodology for real-time joint robotic musicianship.

5.6 Module III: Rhythmic Phrase-Matching Improvisation

The third interaction module that we implemented is a *rhythmic phrase-matching improvisation* module. As in the previous section, this module supports improvisation that is beat- and chord-synchronized to the human player. In addition, it attempts to match the style and density of the human player, and generate improvisational phrases inspired by the human playing.

Beat tracking and chord classification is done as in the opportunistic overlay improvisation. To recap: the timing and pitch of the bass notes are used for detecting the beat, for synchronizing the downbeats of the human's playing, as well as for chord classification.

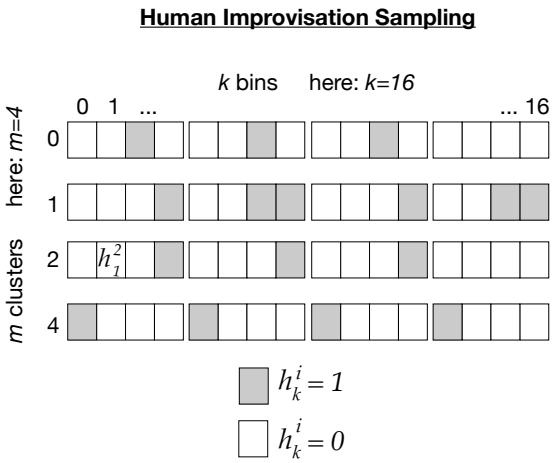

Fig. 11: Clustering of human play vector for Rhythmic Phrase-Matching Improvisation

In addition, this module uses a decaying-history probability distribution to generate improvisational phrases that are rhythm-similar to phrases played by the human. The main gesture of this part selects—in each bar—one of the arm positions that correspond to the currently classified chord. The arm configuration is selected as described in Section 5.4. This is the gesture's anticipatory phase.

When in position, the gesture then plays a rhythmic phrase tempo- and sync-matched to the human's performance. Each arm is separately controlled and plays a different phrase. Arm i plays a phrase based on a probabilistic striking pattern, which can be described as a vector of probabilities

$$\mathbf{p}_i = \{\, p_0^i \; p_1^i \; \cdots \; p_k^i \,\} \tag{6}$$

where k is the number of quantizations made. E.g.—on a 4/4 beat with 1/32 note quantization, $k = 32$. Thus, within each bar, arm i will play at time j with a probability of p_j^i.

This probability is calculated based on the decayed history of the human player's quantized playing patterns. The system listens to the human's last beat's improvisation, quantizes the playing into k bins, and then attempts to cluster the notes in the phrase into the number of arms which the robot will use. This clustering is done on a one-dimensional linear model, using only the note pitch as the clustering variable.

Once the clusters have been assigned, we create a human play vector

$$\mathbf{h}_i = \{h_k^i\} = \begin{cases} 1 \text{ if the human played in cluster } i \text{ at time } k \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

Figure 11 illustrates this concept. In the figure, the human playing is quantized into 16 bins, and each then

clustered into four clusters, corresponding to the four arms of the robot. Each box corresponds to one value of h_k^i, where dark boxes mean that the human played a note in this bin/cluster combination.

The probability p_j^i is then updated inductively as follows:

$$p_j^i = h_j^i \lambda + p_j^i (1 - \lambda) \tag{8}$$

where $0 \leq \lambda \leq 1$ is the decay parameter. The smaller λ is, the more the robot's phrase rhythms will fit the last phrase played by the human. The larger, the more it will "fade" between human phrases, and present a mixture of previous phrases. In performance, this parameter is set heuristically,

The result is an improvisation system which plays phrases influenced by the human playing's rhythm, phrases, and density. For example, if the human plays a chord rhythm, then the vectors $\mathbf{h}_i$ would be identical or near-identical for all clusters, resulting in a robot improvisation that will be close to a chord rhythm. However, there is variance in the robot's playing since it is using the human phrases as a probability basis, therefore changing the pattern that the human plays. Also, since the arm positions change according to the current harmonic lead of the human, and the robot's exploration of the chord space, the phrases will never be a precise copy of the human improvisation but only rhythmically inspired.

Moreover, as the probability vectors mix with data from earlier history, the current playing of the robot is always a combination of all the previous human plays. The precise structure of the robot's memory depends on the value of λ, as stated above.

Another example would be the human playing a 1–3–5 arpeggio twice in one bar. This would be clustered into three clusters, each of which would be assigned to one of the arms of the robot, resulting in a similar arpeggio in the robot's improvisation.

An interesting variation on this system is to re-assign clusters not according to their original note-pitch order. This results in the maintenance of the rhythmic structure of the phrase but not the melodic structure. In the performance described below, we have actually used only two clusters and assigned them to cross-over arms, i.e. cluster 0 to arms 0 and 2 and cluster 1 to arms 1 and 3.

Note that this approach maintains our focus on *gestures* as opposed to note sequences, as the clustering records the human's rhythmic gestures, matching different spatial activity regions to probabilities, which are in turn used by the robot to generate its own improvisation. Importantly—in both improvisation modules—the robot never maintains a note-based representation of the keys it is about to play.

Fig. 12: A live performance of the robot *Shimon* using the gesture-based improvisation system was held on April 17th 2009 in Atlanta, GA, USA.

All three improvisation modules above make extensive use of the separate control of sliders and mallets, and of the physical movement based approach described in this paper. Moreover, they all rely on the safety and expressive regulating layers of the motor controllers described above.

6 Live Performance

We have used the described robot and gesture-based improvisation system in several live performances before a public audience. The first show occurred on April 17 2009 in Atlanta, GA, USA. The performance was part of an evening of computer music and was sold-out to an audience of approximately 160 attendants.

The performance was structured around "Jordu", a Jazz standard by Duke Jordan. The first part was an adaptive and synchronized call-and-response, in which the pianist would prompt the robot with a number of renditions of the piece's opening phrase. The robot detected the correct phrase and, using preparatory gesture responded on beat. A shorter version of this interaction was repeated between each of the subsequent performance segments.

The second phase used the introduction's last detected tempo to play a fixed-progression accompaniment to the human's improvisation. Then the robot started playing in *opportunistic overlay improvisation* taking tempo and chord cues from the human player while repeating an "opening-and-closing" breathing-like gesture, over which the rhythmic improvisation was structured.

The next segment employed *rhytmic phrase-matching improvisation*, in which the robot adapted to the human's tempo, density, style, chord progression, and rhyth-

mic phrases. Our gesture-based approach enabled the robot to adapt in real-time, while maintaining an overall uninterrupted visual motion arc, and the machine seemed to be playing in interactive synchrony with the human player.

An interesting result of this improvisation was a constant back-and-forth inspiration between the human and the robotic player. Since the robot's phrases were similar, but not identical to the human's phrases, the human picked up the variations, in return influencing the robot's next iteration of rhythms. This segment was the longest part of the performance.

Finally, a pre-programmed crescendo finale led to the end-chord, which was an anticipatory call-and-response, resulting in a synchronous end of the performance.

The overall performance lasted just under seven minutes. Video recordings of the performance (Hoffman, 2009) were widely republished by the press and viewed by an additional audience of over 70,000 online viewers.

7 Embodiment in Robotic Musicianship

Beyond *Shimon*'s performative functionality, we also use the robot in our laboratory as a research platform to evaluate core hypotheses of Robotic Musicianship (RM). As mentioned in the Introduction, one of the potential benefits of RM over other computer-supported interactive music systems is the generation of music-related physical and visual cues to aid joint musicianship (Weinberg and Driscoll, 2006b). This could, for example, enable better synchrony through the use of anticipation of the robot's moves on the human's part.

In addition, embodiment in non-musical human-robot interaction has been explored and usually been shown to have a significant effect on social interaction and subjects' reported perception of the robot (Kidd and Breazeal, 2004; Bainbridge et al, 2008). Similarly, a robot musician's physical presence could inspire human musicians to be more engaged in the joint activity. The robot's physical movement could also have choreographic and aesthetic effects on both players and audience. And the acoustic sound produced by the robot could similarly contribute to the enjoyment of the musical performance.

We tested some of these hypotheses in a number of experiments using Shimon as an experimental platform. In this paper, we discuss the effects of physical embodiment and visual contact on two variables: human-robot synchronization and audience appreciation.

8 Evaluation I: Embodiment and Synchronization

In the performance-related segment of this work, we have addressed the mechanisms aiding the robot's synchronization to the human playing, by using preparatory movements and beat and onset detection to play on-beat. The other side of a jointly synchronized performance is the human's ability to coordinate their playing with that of the robot.

Several works have investigated synchronization between humans and robots in a musical or pseudo-musical setting, e.g. (Komatsu and Miyake, 2004; Crick and Scassellati, 2006), however these works have been solely concerned with the synchronization of simple oscillating events, and not with structural and melodic interactions between humans and robots. Moreover, these works only address the robot's synchronization with a human guide, while the work presented here also addresses the reverse issue of the human's synchronizing with the robot's playing.

In line with our view of Robotic Musicianship, we predict that human musicians, when trying to play synchronously with a robot, will take advantage of the visual and physical presence of the machine in order to anticipate the robot's timing, and thus coordinate their playing with that of the robot. However, due to the auditory and rhythmic nature of music, human musicians have also been known to be able to play jointly with no visual cues, and without any physical co-presence. We thus tested to what extent robot embodiment aids in synchronization, and to what extent this effect can be related to the visual connection between the human and the robot.

8.1 Hypotheses

In particular, we tested the following core hypotheses regarding a human musician's ability to synchronize their playing with an artificial (computer or robotic) musician:

H1 — Synchronization is enhanced by the physical presence of a computer musician (Embodiment effect)

H2 — Synchronization is enhanced by visual contact with an embodied computer musician (Visual contact effect)

H3 — The above effects are more pronounced in situations of low accuracy on the part of the computer musician

8.2 Experimental Design

To evaluate these embodiment effects on human-robot musical synchronization, we conducted a 3x2 within-subject study manipulating for level of embodiment and robot accuracy.

Six experienced pianists from the Georgia Tech Music Department were asked to repeat the call-and-response segment from "Jordu" described above, jointly with a robotic musician. The interaction starts by the pianist playing the 7-note introductory phrase on a grand piano. The robot detects the tempo and bar synchronization of the phrase and responds in a rhythmic three-chord pattern on the marimba. The pianists were asked to synchronize a single bass note with each of the robot's chord, as best they could.

Each pianist repeated the call-and-response sequence 90 times. They were asked to play at a variety of tempos, without specifying the precise tempo to play in.

The timing of the human's playing was recorded through a MIDI interface attached to the grand piano, and the robot's playing time was also recorded, both to millisecond precision. MIDI delays between the human and the robot were accounted for.

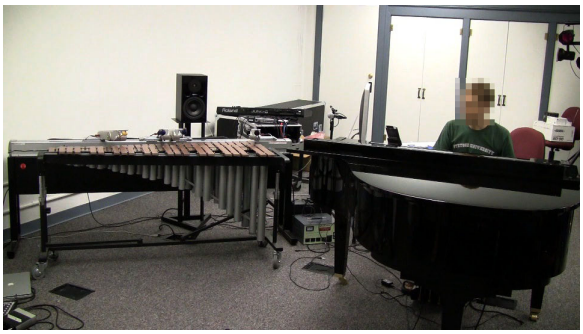

Fig. 13: Experimental setup showing the human pianist on the right, and the robotic marimba player *Shimon* on the left

8.3 Manipulation I: Precision

In the first half of the sequences (the `PRECISE` condition), the robot was programmed to play its response in the precise tempo and on-beat of the human's call phrase. In this condition, the pianists were informed that the robot will try to match their playing precisely.

In second half of the sequences (the `IMPRECISE` condition), the robot was programmed to play its response either on tempo and on-beat to the human's call phrase, slightly slower than the human's introduction phrase (either 50ms too slow, or 100ms too slow), or slightly faster (either 50ms too fast, or 100ms too fast). Note that the first chord is always "on beat" and only the subsequent chords suffer from the accumulative delay.

The pianists were informed that the robot might play slightly off their proposed beat, but that its response will be consistent throughout each individual response sequence. Also, the pianists were asked to try to synchronize their playing with the actual notes of the robot, and not with their own "proposed" tempo and beat.

8.4 Manipulation II: Embodiment

Within each half of the trials—for a third of the interaction sequences (the `VISUAL` condition), the pianists were playing alongside the robot to their right (as shown in Figure 13), enabling visual contact with the robot. In another third of the interaction sequences (the `AUDITORY` condition), the robot is physically present, but separated from the human musician by a screen. In this condition, the human player can hear the robot move and play, but not see it. In the remaining third of the interaction sequences (the `SYNTH` condition), the robot does not move or play. In this condition, the human player can hears a synthesized marimba play over a set of headphones. The order of the conditions was randomized for each subject.

Note that in both the `AUDITORY` and the `SYNTH` condition there is no visual contact with the robot. Furthermore, in both the `VISUAL` and the `AUDITORY` condition there is an acoustic note effect indicating the presence of a physical instrument and a physical player, and in addition, the robot's motor noise can indicate to the pianist that the robot is in motion.[2]

8.5 Results

To account for robot accuracy, and the resulting human uncertainty, we pose three auxiliary hypotheses, differentiating between the three response chords. This is due to the different musical role each chords plays: the first chord occurs an eighth beat after the introductory phrase, so that the pianists can easily synchronize with the robot by simply playing according to their original tempo. The second chord reveals the robot's

[2] For reference, the motor noises peaked at 51.3 dBA measured at a distance of 1.5m length and 1.5m height from the center of the base of the robot. The measurements were made using a calibrated Apex 435 condenser microphone. Measured under the same conditions, without the motors running, the ambient noise in the room was measured at 42.5 dBA.

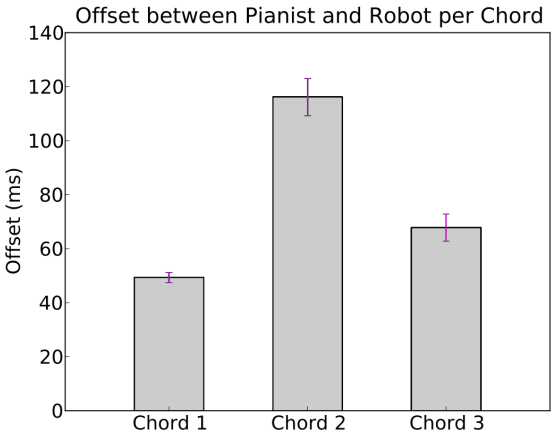

Fig. 14: Mean offset in milliseconds between pianist and robot per chord of the response phrase

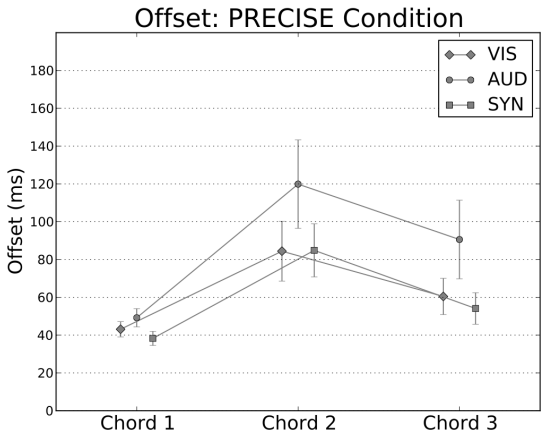

Fig. 15: Mean offset in milliseconds between pianist and robot in the PRECISE condition

perceived tempo, and its temporal placement may vary in the IMPRECISE condition. Since all three chords play at a fixed tempo, the temporal placement of the third chord can be inferred by the interval between the first and the second chord, in which case the synchronization can, again, be achieved by rhythm alone. We thus pose the following auxiliary hypotheses:

H3a — Synchronization in the PRECISE condition is higher than in the IMPRECISE condition
H3b — Synchronization of the first chord is highest
H3c — Synchronization of the second chord is lowest

All three auxiliary hypotheses are supported by our findings.

The offsets in all trials in the PRECISE condition are significantly lower than those in the IMPRECISE condition: 69.63 ± 130.53 vs. 86.91 ± 97.14, T(1513) = -2.89 ***.

Furthermore, as can be see in Figure 14, the offsets (absolute delays) for the first chord are the lowest (49.35ms), those of the second chord are significantly higher (116.16ms), and those for the third chord are lower than the second, but not as low as the first (67.79ms). A one-way ANOVA shows that the three metrics differ significantly (F(2,1512)=47.14, p<0.001***), and pairwise comparison shows that each of them is significantly different from each of the other at p<0.001. We therefore confirm our auxiliary hypotheses **H3a**, **H3b**, and **H3c**, and use these metrics separately in order to evaluate Hypothesis 3.

8.5.1 PRECISE Condition

We first evaluate hypotheses **H1** and **H2** in the PRECISE condition, phrased as follows:

H1a — When the robot is following the human lead precisely, synchronization is enhanced by the physical presence of a computer musician (Embodiment effect with precise robot)
H2a — When the robot is following the human lead precisely, synchronization is enhanced by visual contact with an embodied computer musician (Visual contact effect with precise robot)

Comparing the offset (absolute value of the delay) between pianist and robot in the PRECISE condition, we find a significant difference between the three embodiment conditions using one-way ANOVA [F(2,798)=3.55, $p < 0.05$*] (Figure 15). Post-hoc tests reveal that this is due to the AUDITORY condition being less precise than the other two conditions[T(799)=2.65, $p < 0.01$**]. We thus find no advantage to either visual contact or physical presence with the robot, refuting both Hypothesis **H1a** and **H1b**. This can be attributed to the fact that since the robot plays precisely according to the pianist's cue, the musicians can use their internal rhythm to synchronize with the robot. For possible reasons for the negative effect of the AUDITORY condition, see our discussion below.

8.5.2 IMPRECISE Condition

When the robot changes the detected tempo of the introductory phrase, we expect to detect more of a difference in synchronization between the human pianist and robot. Specifically, we test:

H1b — When the robot is not following the human lead precisely, synchronization is enhanced by the physical presence of a computer musician (Embodiment effect with imprecise robot)

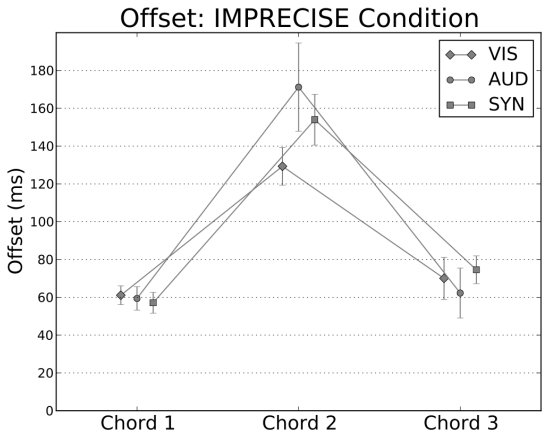

Fig. 16: Mean offset in milliseconds between pianist and robot in the `IMPRECISE` condition

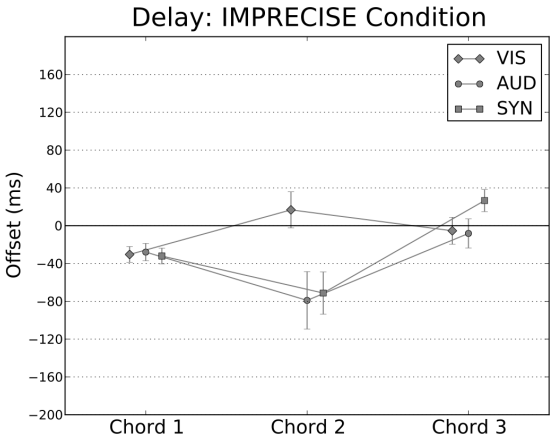

Fig. 17: Mean delay in milliseconds between pianist and robot in the `IMPRECISE` condition

H2b — When the robot is not following the human lead precisely, synchronization is enhanced by visual contact with an embodied computer musician (Visual contact effect with imprecise robot)

H3 — The above effects are more pronounced in situations of uncertainty

Figure 16 shows the mean and standard error for all trials in the `IMPRECISE` condition.

For the first and third chord, we see no difference between the conditions, indicating that, indeed, the human musicians could use the auditory and rhythmic cues to synchronize these two chords. In particular, it is notable that the first two chords are enough for the subjects to synchronize the third chord based on the same interval, supporting Hypothesis **H3**.

However, for the second chord—the timing of which has some uncertainty—the offset is smaller for the `VISUAL` condition compared to both non-visual conditions. This difference is nearly significant: `VISUAL`: 129.32 ± 10.01 ms ; other conditions: 162.80 ± 13.62 ms, [T(182)=-1.66, p=0.09], suggesting that visual cues are used to synchronize the relatively unpredictably-timed event. We did not have access to additional musicians to improve the significance, but this finding encourages additional study in support of Hypothesis **H2b**.

The "offset" discussed above is the absolute error between the human's key-hit and the robot's marimba-strike. The effect of visual contact is, however, more apparent, when looking at the sign of the error: evaluating the directional delay, we find a significant difference between the three conditions on Chord 2 [$F(2,181)=4.80$, $p < 0.01$**]. Specifically, the `VISUAL` condition is significantly different from other two conditions (Figure 17): `VISUAL`: 16.78 ± 19.11 ms; other conditions: -75.21 ± 18.95 ms, [$T(182)=3.10$, $p < 0.01$**]. This finding, too, supports Hypothesis **H2b**.

In particular, we find that trials in the `VISUAL` condition to be delayed with respect to the robot, whereas the trials in the non-visual conditions pre-empt the robot's playing, indicating that pianists *react* to the robot's movement when they can see it, but try to anticipate the robot's timing when they cannot see it.

8.5.3 Effects of tempo

We also find that the benefits of a visual connection increase at slower playing tempos. Figures 18 and 19 show the errors for all trials over and under 100 beats per minutes, respectively, showing a larger embodiment effect for slow trials than for fast trials.

While the `AUDITORY` condition is significantly more error-prone in slow trials than in fast trials (234.54 ± 56.25ms [slow] vs 131.25 ± 10.75ms [fast]; $T(57)=2.14$, $p < 0.05$*), the error in the `VISUAL` condition is not affected by the decrease in tempo (138.59 ± 17.62 ms [slow] vs 119.43 ± 11.54 ms [fast]; $T(60)=0.94$). As above, the effect on the `SYNTH` condition is similar to the `AUDITORY` condition, but less pronounced.

For signed delays, we also find more of the embodiment effect on Chord 2 reported above in slow trials, compared to fast trials (BPM<100: $F(2,69)=4.83$, $p = 0.01$; BPM>100: $F(2,105)=3.11$, $p = 0.048$).

This was an unexpected finding, and calls for further study.

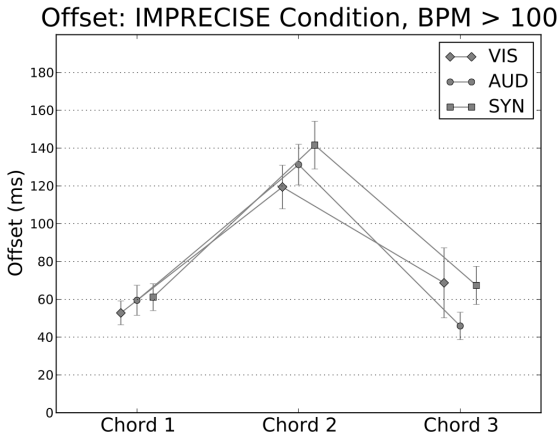

Fig. 18: Mean delay in milliseconds between pianist and robot in the `IMPRECISE` condition for trials over 100 BPM

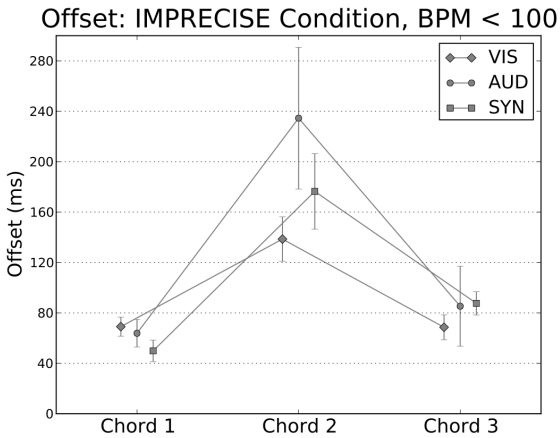

Fig. 19: Mean delay in milliseconds between pianist and robot in the `IMPRECISE` condition for trials under 100 BPM

8.6 Discussion

In our preliminary tests, we find that visual contact with the robot contributes only partially to the degree of synchronization in a call-and-response interaction. The effect of embodiment without visual contact, compared to a disembodied musician, seems to be even less pronounced, and sometimes detrimental.

In the case where the robot does not intentionally change the tempo, we see no advantage to visual contact or embodiment over a synthesized music player. We believe that this is due to the fact that the robot precisely follows the pianist's rhythm, allowing for perfect synchronization simply by playing along in the same rhythm, without any input from the robot.

In the case where the robot slightly alters the tempo in response to the human's playing, we find that the pianists' ability to synchronize with the robot is significantly reduced. For the second chord (the only chord in which there is an uncertainty in timing), visual contact reduces the error compared to the auditory and synthesized condition. In particular, visual contact allows the pianists to *react* to the robot instead of pre-empting the timing of their playing. This indicates that the pianists use the visual cues to time their playing.

By the third chord, the human players seem to be able to achieve a high level of synchronization regardless of the embodiment of the robot. This may indicate that they resort again to a rhythmic cue based on the first two chords.

We also find that visual contact is more crucial during slow trials, possibly suggesting that visual cues are slow to be processed and do not aid much in fast sequences. For example, it may be that during fast sequences, the pianists did not have time to look at the robot. In general, it seems that pianists use visual information when they can, but can resort to rhythmic and auditory cues when necessary and possible.

The limited effect of visual contact could be due to the fact that the expressive characteristics of the robot are somewhat limited, and that the robot does not have specific expressive physical features, such as a head or a face, which could be used for visual coordination. In current work, we have designed and built a head for social communication between the robot and human musicians (see: Section 10). We plan to repeat these experiments with the socially expressive head to evaluate the effects of a dedicated social communication channel to Robotic Musicianship.

Interestingly, it seems that the synthesized condition is less error-prone than the present-but-screened (`AUDITORY`) condition in both precise and imprecise playing modes of the robot. This may be due to the fact that the pianists try to use the existing motor noise from the robot as a synchronization signal, but find it to be unreliable or distracting.

9 Evaluation II: Embodiment and Audience Appreciation

We also tested the effects of visual contact and embodiment on audience appreciation. In this experiment, we filmed two pianists playing in three different improvisation settings each with the robot. We wanted to test how embodiment and visual contact affects joint improvisation as judged by an audience. The physical setup was similar to the previous experiment, and the

conditions were similar to those in the "Embodiment" manipulation, namely `VISUAL`, `AUDITORY` (present but occluded), and `SYNTH`. The only difference was that, in this experiment, the synthesized sound came out of a speaker behind the robot, instead of through headphones.

In this experiment we tested the following hypotheses:

H4 — Visual Contact between a robot and a human musician positively affects audience appreciation of a joint improvisation session

H5 — Physical embodied presence positively affects audience appreciation of a joint improvisation between a human and a machine

9.1 Experimental Setup

The pianists' sessions were videotaped, and from each session, a 30 second clip was extracted by choosing the 30 seconds after the first note that the robot or computer played. We posted these video clips onto a dedicated website, and asked an online audience to rate the clips on eleven scales. Each scale was a statement, such as "The robot played well" (see: Table 1 for all scales), and the subjects were asked to rate their agreement with the statement on a 7-point Likert scale between "Not at all" (1) and "Very much" (7). Subjects watched an introductory clip familiarizing them with the robot, and could play and stop the clip as many times as they liked.

For each pianist, the order of conditions was randomized, but the conditions were grouped for each pianist, to allow for comparison and compensation of each pianist's style. The wording of each scale was matched to the clip, i.e. in the `SYNTH` condition, the word "robot" was replaced by the word "computer".

We collected 30 responses, out of which 21 were valid, in the sense that the subjects rates all three performances of at least one pianist. The reported age of the respondents ranged between 25 and 41, and 58% identified as female.

9.2 Results

In order to compensate for each pianist's style, we evaluated the difference between conditions for each subject and each pianists. We then combined the results for both pianists across all subjects.

Table 1 and Figure 21 show the results of comparing the `VISUAL` condition to the `AUDITORY` condition, and the comparison of the `AUDITORY` condition to the `SYNTH` condition. The first comparison indicates the effect of visual contact between the pianist and the machine, the second comparison indicates the effect of physical co-presence, acoustic sound, and ambient motor noise in the absence of visual contact.

9.2.1 Effects of Visual Contact

We found a significant difference in audience appreciation of the improvisation session between the visual-contact and occluded conditions, on all scales but one (overall enjoyment). Specifically, we find that, even though the robot uses the same improvisation algorithm in all conditions, audiences felt that in the `VISUAL` condition the robot played better, more like a human, was more responsive, and seemed inspired by the human. In addition, we find that the human player, too, was rated as more responsive to the machine, and as more inspired by the robot. The overall rating of the duo as being well synchronized, coordinated, connected, coherent, and communicating was also significantly higher in the `VISUAL` condition.

These findings support hypothesis **H4**, indicating that visual contact between human and robot contributes significantly to the audience's appreciation of robotic musicianship.

9.2.2 Effects of Embodied Presence / Acoustic Sound

In contrast, we could not support hypothesis **H5**, and found only little significant difference in audience appreciation between both occluded conditions. For most scales, there was no difference whether the machine's improvisation came out of the speaker or whether the robot played physically on the keys. The two scales on which there was a significant advantage for the physically embodied/acoustic condition was the robot's responsiveness, and the robot's inspiration by the human player.

Thus, while subjects did not rate the robot's playing as better or more human, they did attribute more human-like characteristics to the acoustically playing robot. Algorithmically, there was no difference between the robot's responsiveness in both conditions, but the robot seemed more responsive and more inspired when it was playing a real acoustic instrument.

Interestingly, occluding the physical robot seems to impair the duo's performance, as in all joint-performance ratings, there is no difference between the occluded physical robot and the synthesized (also occluded) speaker. It is possible that the pianist's engagement drops significantly when there is no visual contact.

Table 1: Effects of visual contact and embodied presence / acoustic sound on audience appreciation of a number of scales. T numbers indicate 1-sample T-Test with $\bar{x} = 0$ as the null hypothesis.
$*p < 0.05; **p < 0.01; ***p < 0.001$

Scale	**VIS-AUD**		**AUD-SYN**	
	$\bar{x} \pm \sigma$	T(38)	$\bar{x} \pm \sigma$	T(38)
I enjoyed this performance	0.28 ± 1.28	1.36	0.26 ± 1.81	0.87
The robot played well	0.92 ± 1.42	4.00 ***	0.23 ± 1.78	0.80
The robot played like a human would	1.11 ± 1.25	5.37 ***	0.05 ± 1.38	0.23
The robot was responsive to the human	0.54 ± 1.28	2.60 *	0.67 ± 1.68	2.44 *
The human was responsive to the robot	1.08 ± 1.79	3.71 ***	-0.28 ± 1.87	-0.93
The duo was well-coordinated and synchronized	1.00 ± 1.48	4.15 ***	-0.28 ± 2.07	-0.84
The human seemed inspired by the robot	1.13 ± 1.90	3.67 ***	-0.26 ± 1.71	-0.93
The robot seemed inspired by the human	0.67 ± 1.37	3.01 **	0.64 ± 1.70	2.32 *
The two players felt connected to each other	0.97 ± 1.58	3.75 ***	0.24 ± 1.81	0.79
The duo felt like a single unit	0.95 ± 1.63	3.58 ***	-0.08 ± 2.06	-0.23
The duo communicated well	1.11 ± 1.74	3.85 ***	-0.05 ± 1.99	-0.16

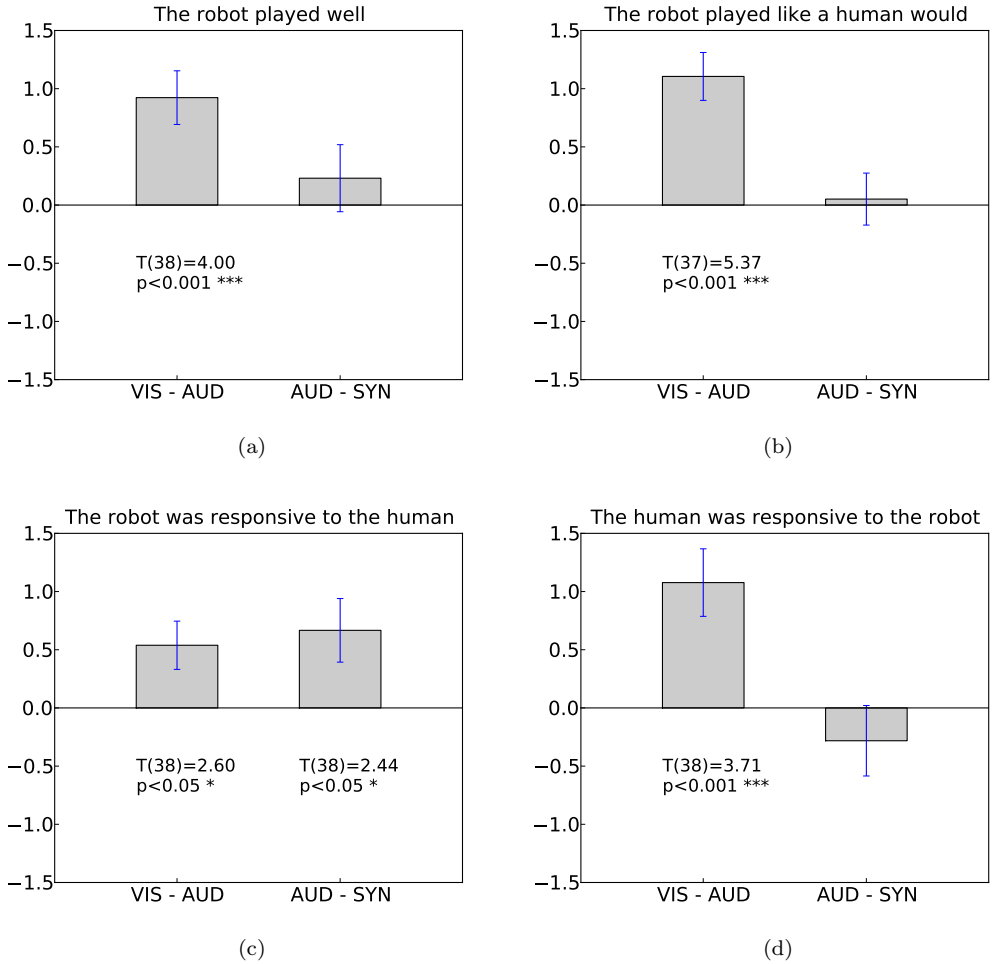

Fig. 20: Effects of visual contact and embodied presence / acoustic sound on audience appreciation of a number of scales. T numbers indicate 1-sample T-Test with $\bar{x} = 0$ as the null hypothesis

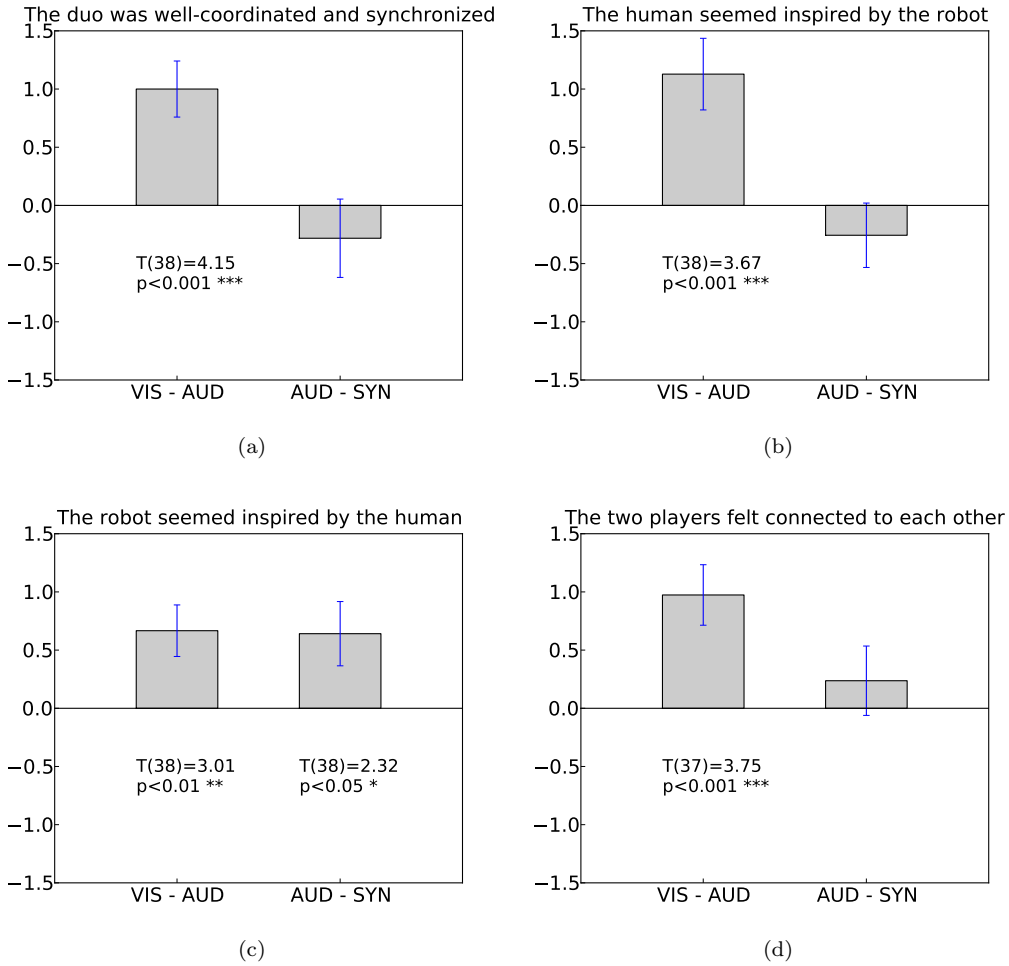

Fig. 21: Cont'd: Effects of visual contact and embodied presence / acoustic sound on audience appreciation of a number of scales. T numbers indicate 1-sample T-Test with $\bar{x} = 0$ as the null hypothesis

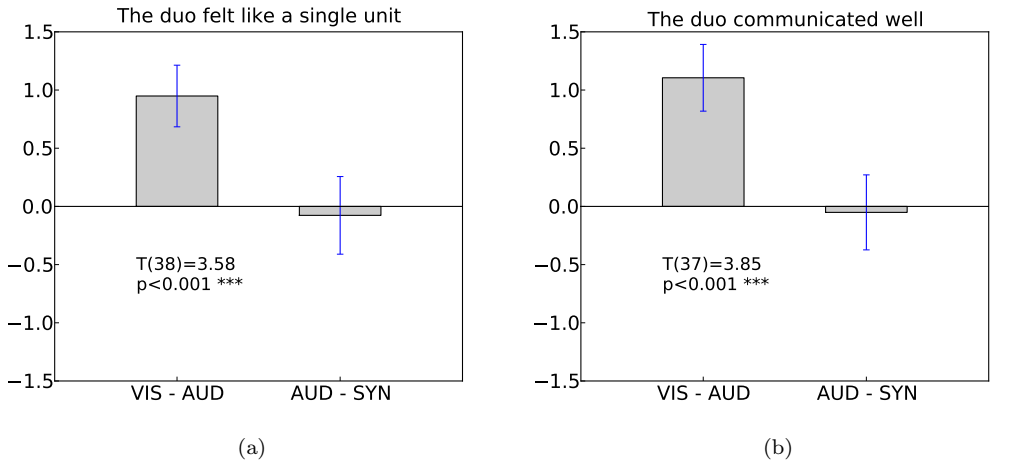

Fig. 22: Cont'd: Effects of visual contact and embodied presence / acoustic sound on audience appreciation of a number of scales. T numbers indicate 1-sample T-Test with $\bar{x} = 0$ as the null hypothesis

We find it surprising that the robot's physical movement and acoustic sound does not contribute to the robot's rated performance quality. However, it is possible that the quality of the video recording and the digital quality of the audio recording are responsible for there being little effect on the audience. This claim should be evaluated in a live-presence human subject study.

10 Future Work

In further development of our robotic improvisation system, we are developing a novel predictive anticipatory system to allow the robot to use past interactions to generate preparatory gestures, based on our findings on anticipatory human-robot interaction (Hoffman and Breazeal, 2007, 2008).

We are also developing a socially expressive robot head to complement the music-playing arms of *Shimon*. This will allow for an additional channel of embodied and gesture-based communication, and also add a visual modality to the robot's perceptual system, through the use of a built-in high definition camera. Through it, the robot will be able to detect social cues of value for musical performance, such as human gestures. In addition, the robotic head will serve as an expressive social modality, for the robot to communicate internal states, such as rhythm, and synchronization to the human performers. Other musical gestures could be used to manage turn-taking and attention between robot and human musicians, highly enriching synchronization and joint musical interaction.

We further plan to extend both human subject studies to a larger subject-base, including both experienced and unexperienced musicians. Furthermore, we will evaluate the inclusion of the socially expressive head, and test its effects on the use of visual contact, as discussed in Section 9. In addition, we will test the effects of embodiment on a simpler interactions than we did in this work, as well as on ones in which the robot takes the lead in the musical sequence.

While *Shimon* does display a number of novel robotic musical interactions, it is obviously still far from matching any human musician capabilities. In particular, the proposed system does not deal with a number of crucial joint improvisation tasks, which we hope to address in future work: the system does not manage turn-taking in a flexible way apart from the transition between improvisation modules. The system is currently fixed to a single Jazz standard piece, and while it is flexible as to the progression of chords, it relies on a fixed set of chords for each performance. Further, since the system uses the human playing as "inspiration", it is not in position to propose completely novel musical phrases. We have found it, though, to be able to surprise human players by the recombination of ideas. Also, *Shimon* is currently limited to playing with a single musician using a MIDI keyboard, as it does not deal with audio analysis.

11 Conclusion

In this paper, we present *Shimon*, an interactive improvisational robotic marimba player developed for research in Robotic Musicianship. We provide technical details of the musically and visually expressive motor-control system, and a gesture- and anticipation-based improvisation system. The design of these systems stems from our belief, that musical performance is as much about visual choreography and visual communication, as it is about tonal music generation. Furthermore, we argue that a physical motion based approach to musical interaction results in a novel methodology for computer-generated improvisation, one that is more appropriate for real-time joint performance between a human and a robot. We have implemented our system on a full human-robot Jazz performance, and performed live with a human pianist in front of a public audience.

In our lab, we use Shimon to empirically study some of the core hypotheses of Robotic Musicianship. In this paper we evaluate the effect of embodiment on human-robot synchronization. We find that visual contact accounts for some of the capability to synchronize to a fixed-rhythm interaction. However, we also find that humans can compensate for lack of visual contact and use rhythmic cues in the case where visual contact is not available. Visual contact is more valuable when the robot errs or changes the interaction tempo. It is also more valuable in slow tempos and delays, suggesting that using visual information in musical interaction is a relatively slow mechanism, or that the human's internal capability to beat-match is more accurate in faster tempos. In addition, our findings indicate that a visually occluded, but present, robot is distracting and does not aid in synchronization, and may even detract from it.

In a study evaluating the effects of embodiment and visual contact on audience appreciation, we find that visual contact in joint Jazz improvisation makes for a performance in which audiences rate the robot as playing better, more like a human, as more responsive, and as more inspired by the human. They also rate the duo as better synchronized, more coherent, communicating, and coordinated; and the human as more inspired and more responsive. There seem to be only a small effect

caused by the acoustic presence of the robot when compared to a synthesized algorithm. That said, an acoustic robot seems more responsive and more inspired. The small effect on other scales could be due to the fact that the study was conducted through video. We plan to extend these preliminary studies to a wider audience, and in particular to also test them with subjects in a live audience, as well as to different joint music scenarios.

References

Baginsky N (2004) The three sirens: a self-learning robotic rock band. http://www.the-three-sirens.info/

Bainbridge W, Hart J, Kim E, Scassellati B (2008) The effect of presence on human-robot interaction. In: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)

Cadoz C, Wanderley MM (2000) Gesture - music. In: Wanderley MM, Battier M (eds) Trends in Gestural Control of Music, Ircam - Centre Pompidou, Paris, France, pp 71–94

Crick C, Scassellati B (2006) Synchronization in social tasks: Robotic drumming. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Reading, UK

Dannenberg RB, Brown B, Zeglin G, Lupish R (2005) Mcblare: a robotic bagpipe player. In: NIME '05: Proceedings of the 2005 conference on New interfaces for musical expression, National University of Singapore, Singapore, Singapore, pp 80–84

Degallier S, Santos C, Righetti L, Ijspeert A (2006) Movement generation using dynamical systems: a humanoid robot performing a drumming task. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS06)

Hoffman G (2009) Human-robot jazz improvisation (full performance). http://www.youtube.com/watch?v=qy02lwvGv3U

Hoffman G, Breazeal C (2004) Collaboration in human-robot teams. In: Proc. of the AIAA 1st Intelligent Systems Technical Conference, AIAA, Chicago, IL, USA

Hoffman G, Breazeal C (2006) Robotic partners' bodies and minds: An embodied approach to fluid human-robot collaboration. In: Fifth International Workshop on Cognitive Robotics, AAAI'06

Hoffman G, Breazeal C (2007) Cost-based anticipatory action-selection for human-robot fluency. IEEE Transactions on Robotics and Automation 23(5):952–961

Hoffman G, Breazeal C (2008) Anticipatory perceptual simulation for human-robot joint practice: Theory and application study. In: Proceedings of the 23rd AAAI Confererence for Artificial Intelligence (AAAI'08)

Hoffman G, Weinberg G (2010) Gesture-based human-robot jazz improvisation. In: Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA)

Hoffman G, Kubat R, Breazeal C (2008) A hybrid control system for puppeterring a live robotic stage actor. In: Proceedings of the17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)

Kidd C, Breazeal C (2004) Effect of a robot on user perceptions. In: Proceedings of theIEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004)

Komatsu T, Miyake Y (2004) Temporal development of dual timing mechanism in synchronization tapping task. In: Proceedings of the13th IEEE International Workshop on Robot and Human Communication (RO-MAN 2004)

Lasseter J (1987) Principles of traditional animation applied to 3d computer animation. Computer Graphics 21(4):35–44

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 10:707

Lim A, Mizumoto T, Cahier L, Otsuka T, Takahashi T, Komatani K, Ogata T, Okuno H (2010) Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pp 1964 – 1969, DOI 10.1109/IROS.2010.5650427

Meisner S, Longwell D (1987) Sanford Meisner on Acting, 1st edn. Vintage

Petersen K, Solis J, Takanishi A (2010) Musical-based interaction system for the waseda flutist robot. Autonomous Robots 28:471–488, URL http://dx.doi.org/10.1007/s10514-010-9180-5, 10.1007/s10514-010-9180-5

Rowe R (2001) Machine musicianship. MIT Press, Cambridge, MA

Singer E, Larke K, Bianciardi D (2003) Lemur guitarbot: Midi robotic string instrument. In: NIME '03: Proceedings of the 2003 conference on New interfaces for musical expression, National University of Singapore, Singapore, Singapore, pp 188–191

Solis J, Taniguchi K, Ninomiya T, Petersen K, Yamamoto T, Takanishi A (2009) Implementation of an auditory feedback control system on an anthro-

pomorphic flutist robot inspired on the performance of a professional flutist. Advanced Robotics 23:1849–1871, DOI doi:10.1163/016918609X12518783330207, URL http://www.ingentaconnect.com/content/vsp/arb/2009/00000023/00000014/art00003

Toyota (2010) Trumpet robot. http://www2.toyota.co.jp/en/tech/robot/p_robot/

Weinberg G, Driscoll S (2006a) Robot-human interaction with an anthropomorphic percussionist. In: Proceedings of International ACM Computer Human Interaction Conference (CHI 2006), Montréal, Canada, pp 1229–1232

Weinberg G, Driscoll S (2006b) Toward robotic musicianship. Computer Music Journal 30(4):28–45

Weinberg G, Driscoll S (2007) The design of a perceptual and improvisational robotic marimba player. In: Proceedings of the18th IEEE Symposium on Robot and Human Interactive Communication (RO-MAN 2007), Jeju, Korea, pp 769–774

Ye P, Kim M, Suzuki K (2010) A robot musician interacting with a human partner through initiative exchange. In: Proc. of 10th Intl. Conf. on New Interfaces for Musical Expression (NIME2010)