A Robot Musician Interacting with a Human Partner through Initiative Exchange

Ye Pan
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Japan
panye@ai.iit.tsukuba.ac.jp

Min-Gyu Kim
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Japan
mingyu@ai.iit.tsukuba.ac.jp

Kenji Suzuki
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Japan
kenji@ieee.org

ABSTRACT

This paper proposes a novel method to realize an initiative exchange for robot. A humanoid robot plays vibraphone exchanging initiative with a human performer by perceiving multimodal cues in real time. It understands the initiative exchange cues through vision and audio information. In order to achieve the natural initiative exchange between a human and a robot in musical performance, we built the system and the software architecture and carried out the experiments for fundamental algorithms which are necessary to the initiative exchange.

Keywords

Human-robot interaction, initiative exchange, prediction

1. INTRODUCTION

When humans make a collaborative musical performance, the initiative exchange often takes place. Sometimes the initiative is exchanged with gestures such as sign and wink and sometimes with a control key which is embedded in the performed musical sounds. It is important to communicate with partner players through the musical sounds without ruining the performance coordination.

Current researches in musical robotics concentrate mainly on playing instruments with precise movements. Few researches address the ability to listen to players' performance, to analyze perceptual musical aspects and to utilize the analysis for playing. The humanoid robot in [2] developed by Toyota Motors can play violin with its dexterous robot hands. Also, the humanoid robot ASIMO in [3] successfully conducted Detroit orchestra showing human-like behaviors. However, in spite of the mechanically high performance, it is still difficult to combine the physical ability with a social dimension which makes it possible for robots to interact naturally with humans and surroundings. There have been some achievements in enabling robots to play a rhythmic sound with sensory perceptions. Nico in [1] and Haile in [5] can detect the stroke timings from a monophonic audio signals of human drumming through embedded auditory sensors and play the drums synchronously with the human player. In [6], Yonezawa *et al* have developed the musical expressive doll to support conversation. But, the application is different from our research because we aim to implement natural interactions in real musical performances.

This research paid attention to the natural interactions between human and robot in terms of the initiative exchange during a musical performance. The objective is to build a robot vibraphone player which is capable of giving and taking the initiative with a human partner by perceiving auditory and visual information. The robot understands the human player's intentions for initiative exchange by listening to the decreased volume and detecting nodding actions. It then performs actions to transfer the initiative with the human player.

For the goal, we have designed the initiative exchange algorithm by linking behavioral cues with auditory signs. We also have realized the fundamental algorithms such as the real-time beat time prediction for synchronization, the human player's nodding detection and the mallet detection to produce natural initiative exchange situation. The cognitive understanding to initiative exchange would create a sense of interaction with a life-like machine and drive the human's affective responses. This is a comprehensive approach that robot imitates the human intelligence of associating the thoughts with the body. From this point of view, this study differs from many previous researches that have focused on imitating the external body motions of humans.

2. INITIATIVE EXCHANGE

The initiative exchange is considered as one of typical characteristics in human-human interaction, which is observed in various situations such as conversation (turn taking), cooperative bodily movements, and collaborative dance or musical performances. In the human-machine musical interaction, the term initiative which is derived from [4] is the authority to vary the performance volume. In this study, we extended the definition of initiative exchange based on auditory cues to multimodal one by associating with nonverbal signs and gestures. The initiative exchange allows more natural turn-taking between human partner and robotic player to play the music cooperatively.

In order to realize a smooth initiative exchange between a human and a machine, the conditions for the occurrence of the initiative exchange must be selected to suite the human intuitive senses. We considered the following assumptions for initiative exchange.

A. Both the human performer and the robot synchronize with mutual solo performance to show that s/he is ready to give/take the initiative. If they stop playing vibraphone, it is assumed that they do not intend to take the initiative.

B. When the human player increases/decreases performance volume beyond a particular threshold value,

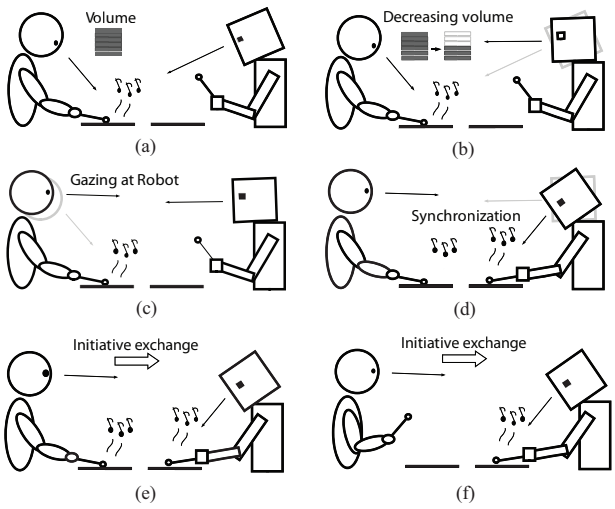

Figure 1: The process of how initiative is transferred from the human player to the robot player

the robot player looks at the human player because it is the critical state which represents that the initiative exchange is ready to occur.

C. The human player's nodding confirms demand or abandonment for the initiative exchange.

Based on the above assumptions, we will describe how the initiative can be transferred between a human performer and a robot player. The beat synchronization in cooperative music playing is considered as the beginning sign of interaction between the human performer and the robotic player.

Figure 1 illustrates the initiative exchange manner from the human to the robot. In the beginning of the musical performance, the human performer plays vibraphone solo. The robot looks at the human performer's vibraphone and listens to the music played by the human player (See Figure 1(a)). After a while, the human performer starts to decrease the performance volume in order to show his/her intention to give the initiative. If the performance volume is decreased below the predetermined threshold value, the robot detects the cue to get the initiative and look up at human performer (See Figure 1(b)). The robot then recognizes the human performer's head orientation and nodding. They indicates that s/he wants to transfer the initiative to the robot (See Figure 1(c)). After that, the robot tries to play vibraphone to synchronize the beat time with the human performer to take the initiative from the human performer (See Figure 1(d)). Despite the robot received the initiative, if the human performer continues to play with robot, it is regarded as the sign that human player still wants to play a supporting role in the music performance (See Figure 1(e)). On the other hand, if the human performer stops playing, the initiative is transferred to the robot player (See Figure 1(f)). When the robot player takes the initiative, there will be two possible cases as described in Figure 1(e) and 1(f).

1) The human performer plays with robotic player.

2) The robot player plays vibraphone solo.

In case 1, if the human player begins to increase the volume, the robot lifts up the head to look at the human player while continuing to play. In this situation, if the human gives a nod, s/he intends to play together. The increasing volume and the nodding represent the cues for playing together. On the other hand, for the case 2, if the human player stops to play after a while, the robot can take the initiative totally and then it plays vibraphone solo as mentioned above.

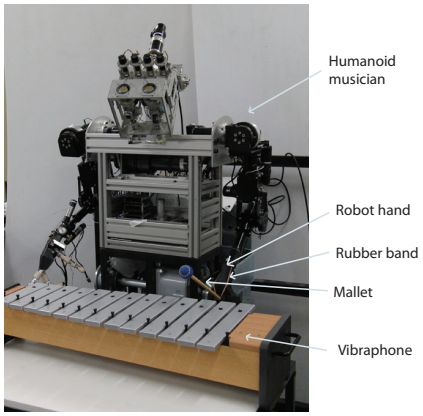

Figure 2: The humanoid musician

In this research, the initiative exchange is inherent in the musical performance without producing voices or using external devices. The specific gestures such as looking and nodding are only used for the initiative exchange with auditory cues such as increasing/decreasing the performance volume. In this research, we did not use the internal states as the template for the robot performance.

3. SYSTEM OVERVIEW

The overall system consists of three different kinds of servers: the auditory, the vision, and the robot control sever. Each sever communicates through TCP/IP. In the auditory sever, the sound processing algorithms were implemented with Max/MSP on Windows XP. It mainly plays the role of analyzing the sound played by both of the human and the robot. The beat time prediction algorithm for synchronization is also carried out in the auditory server. In the vision server, the gesture recognition to detect head orientation and nodding was developed with OpenCV based on C++. Regarding the robot control server, the position control for playing vibraphone was achieved based on a real-time Linux system (ART-Linux). Figure 2 shows the humanoid robot and the designed simple robot hand. The robot has two arms, a torso, and a head with a vision camera. It has 7 DOF arm and multi-fingered robot hand on the right side. The left arm is 5 DOF: 2 DOF shoulder (pitch and yaw), 2 DOF elbow (pitch and roll) and 1 DOF wrist (rolling). For playing the vibraphone, we used the left elbow's pitching motion. We designed a simple robot hand which has a mallet holder and a rubber band. The mallet holder allows the mallet to move only in the vertical direction. In order to let the mallet easily return to the initial position after hitting the vibraphone, the rubber band was used.

4. METHODOLOGY

In this section, we will explain the detail functions of the audio processing module, the vision processing module and the robot control module which are included in software architecture shown in Figure 3.

4.1 Audio Processing Module

The audio processing module is mainly used to analyze audio input signals from the microphone which are the play-

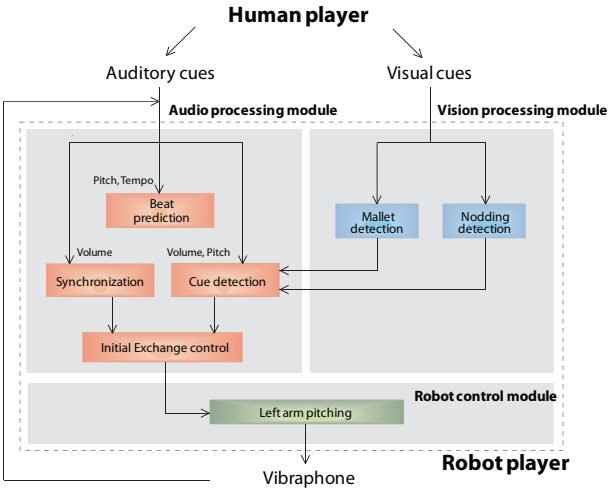

Figure 3: Software architecture

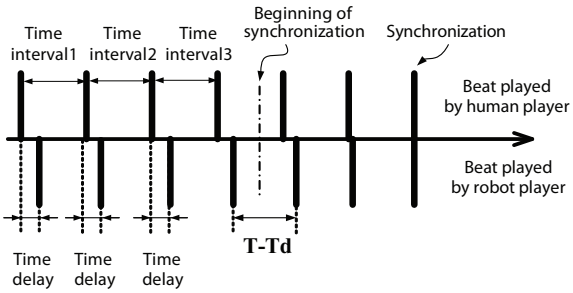

Figure 4: Prediction algorithm

ing sounds by the human and the robot player. The audio processing module includes three main functions as below;

(I) Beat time prediction for synchronization: The audio processing module distinguishes the human's and robot's playing sound based on the loudness and the performance pitch. This module estimates the next time which the robot player should beat for synchronization with the simple beat time prediction algorithm as depicted in Figure 4.

Given a monotonic score, the human and the robot players play different tones in the vibraphone. Therefore, the performance pitch information can be used to distinguish the players' sounds. The performance tempo **T** is calculated by collecting the latest three time intervals of the human's playing sounds. Before the synchronization, it is supposed that the time delay of the robot player is a constant **Td**. The assumption was verified through an experiment. After the robot begins to synchronize the beat time, the beat time prediction makes **T-Td** to approximate T and then sends the swing command to the robot control server.

(II) Cue recognition of the human player: The performance volume change of the human performer's playing is regarded as the cue which leads the robot player to look at the human performer.

(III) Conveying the arm swing command: We defined an attack as the amplitude of the player's sound which exceeds the predetermined threshold value. We decided the pitching angle of the left arm from the difference between the lifting-up position and the striking position. If the attack from the human performer is

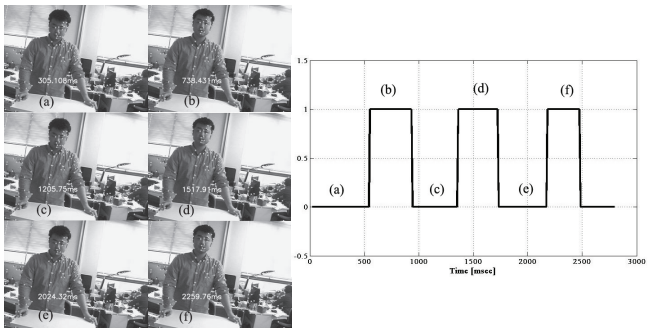

Figure 5: Nodding detection

Figure 6: Mallet detection

detected, then the command of the pitching action is sent to the robot player. Otherwise, the robot player remains stand by until the next command is received.

4.2 Vision Processing Module

The robot musician has a CCD camera mounted on the head. The vision processing module is used to recognize the gestural signs from the human performer and to detect the human player's and the robot player's mallet positions.

(I) Nodding Detection: We adopted an optical flow approach to estimate object's motion across a series of acquired image frames. Color-based skin detection and the Lucas-Kanade method were implemented for the nodding detection. First of all, the random points are distributed on the initial image to track the changes in the next images. When the human player nods his head, position of the points on the face will be conspicuously changed. Figure 5 shows an example of nodding detection.

(II) Mallet Detection: Mean-shift is the method used to track the mallet's endpoint. Initially given the color of the mallet endpoint, the vision processing module tracks the endpoint location. The location can be used for the robot player to revise its playing in real time with the auditory feedback. Figure 6 shows the results of tracking the mallet endpoint.

4.3 Robot Control Module

First of all, in the initial state the robot player waits for starting the music performance with fully extending down its arm. In order to begin to play, the robot raises up the arm over vibraphone as the lifting-arm state. When the robot player tries to hit the vibraphone, it extends the arm by moving its elbow until the robot player's mallet almost comes in contact with the vibraphone. Then the robot can attack the vibraphone by means of the inertial moment which is generated by the mallet swing. Likewise, the robot can play vibraphone through raising and extending the arm repeatedly each time. The movement of the robot player was accomplished based on position control.

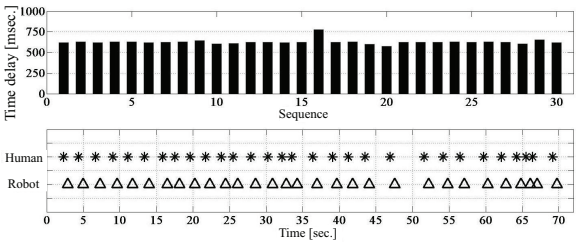

Figure 7: Time delay

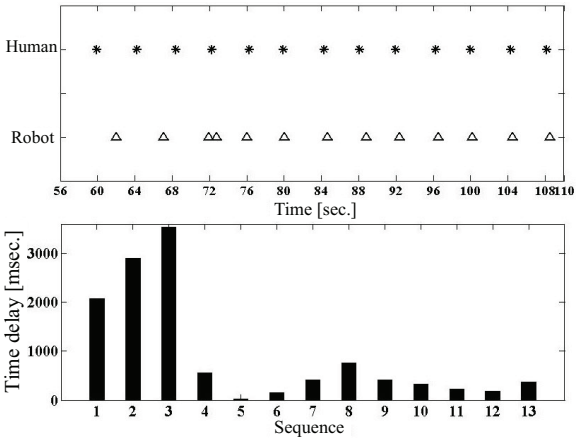

Figure 8: Synchronization at 30 BPM (0-48 sec)

Compared with a human playing motion which is more efficient with very high number of degree of freedom, the robot player's motion is simple but very appropriate for the task.

5. EXPERIMENTS

5.1 Measuring of Time Delay

Figure 7 indicates the measured time delay between the human and the robot playing before the synchronization. The result is the response of the robot player without applying the prediction mechanism. In this experiment, the average tempo of the human performance was 30 BPM (beats per minute). During the 30 times playing, a time delay which has regularly occured was approximately 620 ms in average.

5.2 Synchronization

Considering the measurement of the time delay, it takes about 620 ms from receiving the command to hitting the vibraphone. Figure 8 and Figure 9 show the experiment results of the proposed prediction algorithm. The experiments were carried out for 180 sec. In the first 60 sec, the average tempo of human player's performance was 30 BPM. Then in the next 60 sec, the tempo was changed to 15 BPM and turned back again to 30 BPM in the last 60 sec. The results show that only about 5 beats were needed to synchronize with the human performer.

Since human generally recognizes the difference between the beats under time interval of 100 ms, the robot player did not need to synchronize precisely with the human performer. We have verified that robot player can synchronize with the human performer within a time delay of less than 500 ms. In this research, the maximum time delay 500 ms during the synchronization is acceptable because the robot musician was built up not to play precisely but to evaluate the interaction for the initiative exchange.

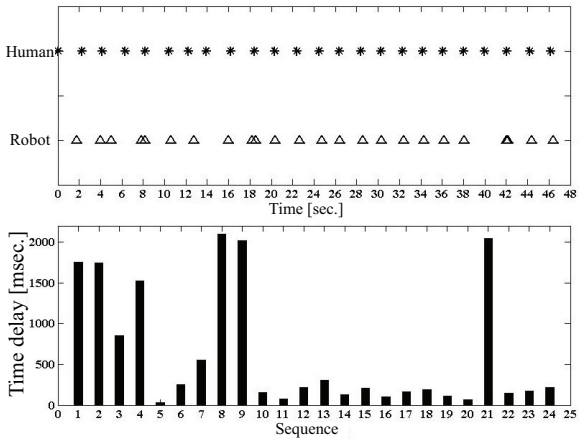

Figure 9: Synchronization at 15 BPM (56-110 sec)

6. CONCLUSION AND FUTURE WORKS

In this paper, we considered that the coordination of nonverbal cues and physical behaviors will be important prerequisite for a comfortable and natural social interaction between a human and a robot as we know that it occurs between human and human.

So, we introduced a new methodology for the initiative exchange by multimodal perception and proposed a foundamental algorithms for the initiative exchange between the robot and the human players. The robot player can understand the cue by looking at the human performer after perceiving the performance volume change. Also, it can recognize the human player's nodding and mallet positions through the vision camera. We finally evaluated the performance of the implemented algorithms through experiments.

In order to play polyphonic sounds, we will improve the motions of the robot player. The gazing direction recognition will be considered to understand human player's intentions better. Also, we will combine the each algorithm to realize the initiative exchange during real music performance and evaluate human's satisfaction while performing with the robot musician.

7. REFERENCES

[1] C. Crick, M. Munz, T. Nad, and B. Scassellati. Robotic drumming: Synchronization in social tasks. *Proc. of the 15th IEEE Int'l Symp. on Robot and Human Interactive Communication*, pages 97–102, 2006.

[2] Y. Kusuda. Toyota's violing playing robot. *Int'l Jour. of Industrial Robot*, 35:504–506, 2008.

[3] H. Masato and T. Tooru. Development of humanoid robot asimo. *Honda R&D Tech. Rev.*, 13(1):1–6, 2001.

[4] Y. Taki, K. Suzuki, and S. Hashimoto. Real-time initiative exchange algorithm for interactive music system. *Proc. of Int'l Conf. of Computer Music*, pages 539–542, 2000.

[5] G. Weinberg and S. Driscoll. Robot-human interaction with an anthropomorphic percussionist. *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, pages 1229–1232, 2006.

[6] T. Yonezawa and K. Mase. Musically expressive doll in face-to-face communication. *Proc. of the 4th IEEE Int'l Conf. of Multimodal Interfaces*, pages 417–422, 2002.