The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Implementation of a Musical Performance Interaction System for the Waseda Flutist Robot: Combining Visual and Acoustic Sensor Input based on Sequential Bayesian Filtering

Klaus Petersen, Student Member IEEE, Jorge Solis, Member IEEE, and Atsuo Takanishi, Member IEEE

Abstract— **The flutist robot WF-4RIV at Waseda University is able to play the flute at the level of an intermediate human player. So far the robot has been able to play in a statically sequenced duet with another musician, individually communicating only by keeping eye-contact. To extend the interactive capabilities of the flutist robot, we have in previous publications described the implementation of a Music-based Interaction System (MbIS). The purpose of this system is to combine information from the robot's visual and aural sensor input signal processing systems to enable musical communication with a partner musician. In this paper we focus on that part of the MbIS that is responsible for mapping the information from the sensor processing system to generate meaningful modulation of the musical output of the robot. We propose a two skill level approach to enable musicians of different ability levels to interact with the robot. When interacting with the flutist robot the device's physical capabilities / limitations need to be taken into account. In the beginner level interaction system the user's input to the robot is filtered in order to adjust it to the state of the robot's breathing system. The advanced level stage uses both the aural and visual sensor processing information. In a teaching phase the musician teaches the robot a tone sequence (by actually performing the sequence) that he relates to a certain instrument movement. In a performance phase, the musician can trigger these taught sequences by performing the according movements. Experiments to validate the functionality of the MbIS approach have been performed and the results are presented in this paper.**

I. INTRODUCTION

At the center of our research at Waseda University is the development of humanoid musical performance robots. These are robots that are able to perform on a musical instrument, accurately emulating the human way of playing. The Waseda Flutist Robot WF-4RIV, has been developed over more than 15 years in several generations. The Waseda Flutist robot is a humanoid robot, it has artificial lungs, an artificial oral cavity, lips and vocal chord, as well as arms and fingers to play a real flute instrument. It has a total of 41-DOFs and is controlled by several computing units, that besides doing the basic motor control also perform visual and aural sensor processing, as well as the translation of sensor information into musical parameter modulation data. Recently research on the robot has progressed to the degree that the robot is able to play the flute at the level of an intermediate human player. The current state of the development of the robot mechanism and control system has been published in [1] and [2].

Regarding this purpose our goal is to have the robot play interactively together with a human band. Musicians in a human Jazz band rely on giving visual and acoustic cues to each other to determine to synchronize each other's play during a performance ([3]). In our research we try to make the robot to be able to react to these cues in a musically meaningful way.

In an approach similar to our work, Gil Weinberg constructed the percussion robot *Haile* ([4]) that is able to tune in to the rhythm of a partner musician and within a certain limit vary its performance to display improvisation capability. One of the main differences to our system are shape and complexity of the robot. Drum playing being a task where only the musicians hand is involved to directly trigger a sound, is compared to playing the flute a relatively simple activity. Weinberg has concentrated on the interaction between human musicians and his music robot. His robot can actively adjust to the play of partner musicians, imitating their behavior of creating a rhythm. His work uses the approach of analyzing the recorded music data and extract information about the current musical situation by applying a rhythmic rule-set. Although we also base our aural interaction on imitating the human musician, we do so by comparing musical input with prerecorded sequences in the library using a histogram method. A further substantial difference is that our system also involves visual processing.

To realize interaction between the robot and human musicians; in [5], we have introduced the Musical based Interaction System (MbIS). So far we have mainly focused on the visual and acoustic sensor processing. In this paper we want to concentrate specifically on the last stage of the robot's performance system for translating sensor data into musically meaningful performance modulation parameters (in the following, this part of the system will be referred to as the mapping module). Our system consists of two levels, one stage for interacting with players of a beginner skill level and one stage for more advanced players.

In the beginner level interaction stage we focus on enabling a user who does not have much experience in com-

Klaus Petersen is with Waseda University, Graduate School of Advanced Science and Engineering, Waseda University, 3-4-1 Ookubo, Shinjuku-ku, 169-8555 Tokyo, Japan (phone: +81-3-5286-3257; fax:+81-3-5273-2209; e-mail: klaus@moegi.waseda.jp).

Jorge Solis is with the Faculty of Science and Engineering, Waseda University and researcher at the Humanoid Robotics Institute (HRI), Waseda University, Tokyo, Japan (e-mail: solis@ieee.org).

Atsuo Takanishi is with the Faculty of Science and Engineering, Waseda University and one of the core members of the Humanoid Robotics Institute, Waseda University, Tokyo, Japan (e-mail: contact@takanishi.mech.waseda.ac.jp).

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

municating with the robot to understand about the device's physical limitations. We use a simple visual controller that has a fixed correlation regarding which performance parameter of the robot it modulates, in order to make this level suitable for beginner players. The WF-4RIV is built with the intention of emulating the parts of the human body that are necessary to play the flute. Therefore it has artificial lungs with a limited volume. Also other sound modulation parameters like the vibrato frequency (generated by an artificial vocal chord) have a certain dynamic range in which they operate. To account for these characteristics the user's input to the robot via the sensor system has to be modified in a way that it does not violate the physical limits of the robot. To modulate the robot's performance parameters we use a motion tracking algorithm to detect a partner musician's instrument movements. For this purpose we introduced specialized controllers (*Virtual Faders* and *Virtual Buttons*) in [6].

In the advanced level interaction interface, our goal is to give the user the possibility to interact with the robot more freely (compared to the beginner level). To achieve this we propose a teaching system that allows the user to link instrument gestures with musical patterns. Here, the correlation of sensor input to sensor output is not fixed. Furthermore, we allow for more degrees-of-freedom in the instrument movements of the user. As a result this level is more suitable for advanced level players. We use the robot's instrument gesture detection system that we have presented previously in [7]. A Bayesian mapping algorithm is employed in order to ensure, that if the teaching musician does not account for all combinations of instrument orientation and musical output in the teaching phase, in the performance phase the robot will automatically play the most closely matching answer modulation to a given instrument state.

Regarding previously published work, learning-teaching techniques similar to the method proposed here have, in various ways, been introduced in robot control. In [8] oral expression and sensory inputs are mapped to control the motor of a robotic arm. The approach uses information gained from camera images and microphone input to set up Hidden Markov Models (HMMs). These models contain a state-space representation of the generated training data. This information is used to execute tasks that consist of combinations of the different teaching situations. Specifically, our approach is adapted from a problem setting, in which a robotic arm is taught how to empty a glass of water into a sink ([9]). The robot learns this movement by demonstration, taking the glass from a specific location. The goal is that, with changing initial location, the robot autonomously is able to find the right way to empty the glass without assistance. Similar to the other approaches referenced above this application uses a state-space table to record the instructions during the teaching phase. During the execution, sequential Bayesian filtering is employed to adapt the learnt data to the current problem environment.

Our new approach brings two significant novelties compared to how we have dealt with sensor input in previous work. First, we now work with the acquired sensor data conditionally by getting feedback from the body of the robot (e.g. state of the lung) and using this feedback to modulate the influence of the user interaction on the performance parameters. Second, in the advanced interaction level, we allow the user to teach the robot how one or more sensor values modify one or more performance parameter values.

II. IMPLEMENTATION OF A MAPPING MODULE FOR THE MBIS

A. *Direct Translation of Sensor Input to Musical Performance Parameters with Consideration of Physical Restraints (Beginner Level Interaction Mapping)*

The purpose of the beginner level interaction system mapping module is to translate the actions of the user that are recorded through the virtual buttons and faders into musical output. This output is to make musical sense in the way that the user can express himself as freely as possible, while at the same time respecting the physical limitations of the robot. One important limitation of the WF-4RIV flutist robot is the restricted air volume that can be contained by the lung. Similar to the human breathing the robot is only able to produce sound for a certain duration, until the lung is empty. The robot has also further limitations, like a maximum playing speed and maximum modulation speed performance parameters like the vibrato frequency.

When receiving data about the robot's partner musician from the vision processing system, we can map this data directly onto a musical performance parameter. In case of receiving a continuous value from a virtual fader controller, this relationship can be formulated as shown below:

$$A(t) = k * I(t) \tag{1}$$

This equation contains the constant k representing a scaling factor to resize the sensor (virtual fader) value $I(t)$ to an appropriate output value A. Using information about the maximum and minimum value emitted from this controller, we can condition k accordingly, so that the output value A does not exceed the acceptable range for the performance parameter.

$$A(t) = k(t) * I(t) \tag{2}$$

with

$$k(t) = \begin{cases} k & \text{if } t < T_{Breathing} \\ 0 & \text{if } t \geq T_{Breathing} \end{cases} \tag{3}$$

Some limitations of the robot however are not time-constant. The capability of the robot to create an air-beam in order to play the flute, depends on the air volume left in the lung. Taking this into account we add time-dependence to k. $T_{Breathing}$ indicates the time duration of air remaining in the lung, enabling a tone to be be produced. The equation expresses that the intended output of the flute robot is to be conditioned with the fill-status of the lung. If the lung becomes empty, the equation constant $k(t)$ is set to 0, resulting in $A(t)$ to become 0 as well. As result the flute robot does no produce sound output.

As the switch from a normal performance to a breathing break is very abrupt, this method might not be satisfactory in a musical sense. Musical progressions are normally characterized by smooth transitions or intentionally inserted breaks at certain points. If a musician needs to interrupt his performance as a result of physical constraints, that would, under normal circumstances, give the impression of an unsatisfactory presentation to the audience. As every human has various bodily constraints, these need to be integrated in the mode of performance in a way, that is as little as possible perceivable by the audience. To address this issue in a natural way we applied a method to provide smoother outline to the switching edges. Using a digital low-pass filter on the time-dependent conditioning as indicated in equation (2), we get smoother outlines for the switching of the performance states (normal play / interrupted play due to lung-refill).

The modulation envelope that results from this method of smoothing is similar to ADSR (Attack-Decay-Sustain-Release) curves used in electronic music synthesizers. If we vary the parameters of the low-pass filter we can change the slope of the attack curve. This enables us to adjust the smoothing in a more human-like fashion. We implemented the digital low-pass filter as a FIR (Finite Impulse Response) filter. This filter achieves a similar effect to value-averaging by chaining delay stages and scaling stages. Considering a queue of values (in this case the time-dependent values of $k(t)$), fractions of previous values are fed-back to the current value and added. Depending on the number of these delay-elements the low-pass / averaging effect becomes stronger.

Through the implementation of this method in the mapping module of the beginner level interaction system we can guarantee that the robot's partner musician can control the robot safely (within the robots value constraints). Using the time-dependent scaling parameter and the smoothing filter, the robot will automatically adjust the sensor input from the vision system to account for the system's mechanical properties (Figure 2).

Although the above principle was so far introduced using a virtual fader as controller source, it also applies to the adjustment of data from a virtual button. In an interaction setting using virtual buttons, the user might be able to switch between various melody patterns. If during the performance of one melody pattern the volume drops due to the necessity of a lung-refill, this might seem unnatural to the audience. A resolution to this problem is to calculate, if the melody pattern to be played fits into the time remaining until the next lung-refill; And in case it does not apply the proposed fade-out effect to the sequence tempo (slow the sequence down) using the low-pass filter to smoothly finish the last note of the pattern still fitting into the lung cycle (Equation (4), (5)).

$$A(t) = k(t) * I(t) \tag{4}$$

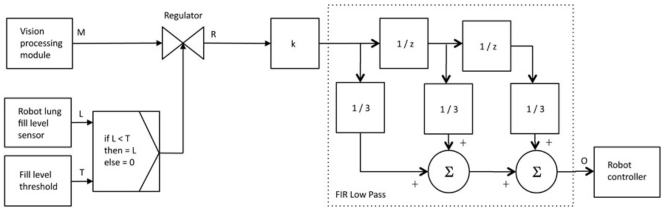

Fig. 1. Block diagram of the beginner level mapping method. M denotes the movement value output from the sensor processing module, L the detected lung fill level, T the fill level threshold, R the movement data regulated by the fill level controller and O the filtered musical parameter modulation output.

with

$$k(t) = \begin{cases} fadeout & \text{IF sufficient residual air volume} \\ play & \text{IF residual air volume not sufficient} \end{cases} \tag{5}$$

B. Translation of Sensor Input to Musical Performance Parameters based on a Sequential Bayesian Filtering Approach (Advanced Level Interaction Mapping)

The mapping approach for the advanced level interaction system, is based on the assumption that the partner musician of the robot is a player of advanced skill level. As a result the method leaves more space for free control of the robot.

The goal of our approach is to implement this technique into our musical interaction setup to create sensible musical output. In contrast to the previous approach, this time we do not use the virtual faders and buttons as input source, but the particle filter-based tracker. In two phases, the teaching phase and the performance phase we try to enable the robot to estimate the song state according to the input received from the vision and audio processing system. If the robot knows the current state of the song, it will be able to play an appropriate reaction to the human partner musician's actions. A deliberate number of input parameters (e.g. horizontal and vertical instrument orientation), is to be mapped to a deliberate number of output parameters (e.g. vibrato amplitude, played note value). This should be done without the teaching musician having to account for all possible state combinations. Using a particle filter, even if during the performance an unknown state combination is given to the robot, it is to automatically play the most closely matching answer modulation.

At first, in the teaching phase, the teacher fills up the state-space table with information on how to relate instrument orientation changes to performance modulation. Although the instrumentalist may spend a long time teaching, this information will probably not be complete. That means that there are states of the instrument configuration that are not accounted for in the table. In the performance phase, the robot reacts to the movements of the musician in order to reproduce the previously learnt behavior. To relate a configuration of the instrument (orientation) to a correct modulation, the robot uses a particle filter (Bayesian filter). In Fig. 2, the robot takes the data from the vision processing, seeds particles in the state-space table (e.g. in the button

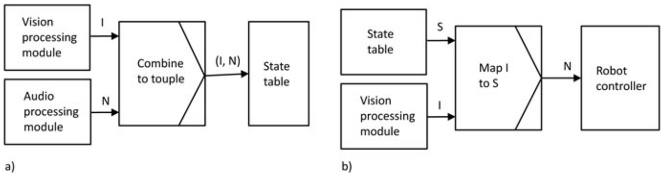

Fig. 2. Block diagram of the advanced level mapping method. a) shows the learning phase signal flow. I denotes the detected instrument motion, N the detected note or rhythm sequence. b) displays the signal flow in the performance phase. Additionally S denotes a state from the state table here (could be also expressed as an (I, N) tuple).

states and fader states column) and selects the most closely related particle. The modulation that this particle relates to (in the table), is played by the flutist robot.

A Bayesian filter represents the PDF (Probability Density Function) $p(\underline{x}_k|\underline{z}_{k-1})$ of state $\underline{x}_k$ given observation $\underline{z}_{k-1}$ were k is the discretized time. Specifically for the particle filter this PDF is defined through a set of N_s random measurements $\underline{s}_k^i$ with weights π_k^i. In this case, the current observation $\underline{X}_k$ is given by

$$\underline{X}_k = \sum_{i=1}^{N_s} \pi_k^i \underline{s}_k^i \tag{6}$$

and the PDF $p(\underline{x}_k|\underline{z}_{k-1})$ can be approximated as ([10])

$$p(\underline{x}_k|\underline{z}_{k-1}) \approx \sum_{i=1}^{N_s} \pi_k^i \delta(\underline{x}_k - \underline{s}_k^i) \tag{7}$$

with

$$\sum_{i=1}^{N_s} \pi_k^i = 1 \tag{8}$$

III. EXPERIMENTS AND RESULTS

Experiments were performed to evaluate, how well a user / musical partner can express his musical intentions using the proposed two stage mapping approach.

In case of the beginner level interaction interface experiment, the robot is controlled by one virtual fader. This fader is used to continuously control the speed of a pre-defined sequence that is played by the flutist robot. The output of the sensor processing system determining the value of the virtual fader is conditioned by the lung movement of the robot. We use the method introduced in section II to continuously reduce the speed of the performed pattern, when we reach a certain fill-level of the robot's lung. In order to perform the experiment, a professional flutist player is situated in front of the robot (within the viewing angle of the robot's cameras). After introducing the functionality of the beginner level stage to the player we recorded data of the resulting interaction with the robot.

To achieve quantitative results for the first level interaction system we performed the experiment with a professional flutist player. A graph of these results is shown in Fig. 3. Fader movements control the tempo of the tone sequence that is performed by the robot. If the amount of air remaining in the lung reaches a certain limit (in this experiment approx. 15% of the lung volume), the fader value transmitted to the robot is faded-out (using the low-pass filter previously described). At $5s$ and $15s$ the robot refills its lungs for a duration of approx. $5s$. These breathing points have a time-distance of approx. $10s$. During the breathing points no sound is produced by the robot. The fader value actually transmitted to the robot is faded out before the lung is completely empty. This adjustment can be observed at $3.5s$-$5s$ and $13.5s$-$15s$ in the fader value plot, the filtered fader value plot and the robot output volume plot. As the fader value is faded-out rapidly, the resulting performance tempo of the robot decreases quickly. In this experiment the robot continuously plays a pattern of the notes $a4$ - $b4$ - $c5$ - $b4$.

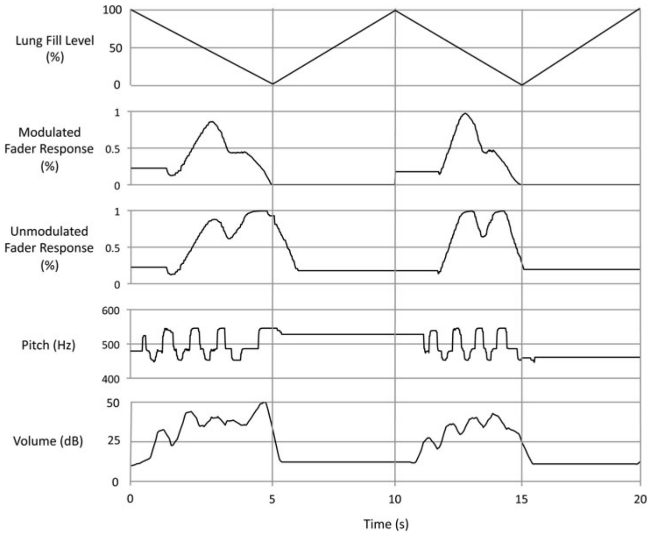

Fig. 3. In the beginner level interaction system the user controls the tempo of a pattern performed by the robot. The lung fill level plotted in the top graph, modulates the input data from the virtual fader resulting in the robot performance displayed by the pitch and the amplitude curve.

In the experiment for the advanced level interaction system we try to confirm that, using our mapping module, a musician of advanced skill level has the possibility to teach the robot how to relate certain instrument movements with the variation of certain musical parameters.

The experiment has two phases, the teaching phase and the performance phase. In the first phase the interacting musician teaches a movement-performance parameter relationship to the robot. In this particular case we relate one of three melody patterns to the inclination angle of the instrument of the robot's partner musician. From this information the robot builds a state-space table that relates instrument angles to musical patterns. In the second stage the interaction partner controls the robot with these movements. Using the proposed particle Bayesian filtering we search for the instrument angle in the state-space table that most likely represents the current state. When a match is determined, the robot plays the musical pattern that relates to the current instrument angle. The transition of the teaching phase to the performance phase is defined by the number of melody patterns associated by the robot. In case of this experiment, the switch occurs after

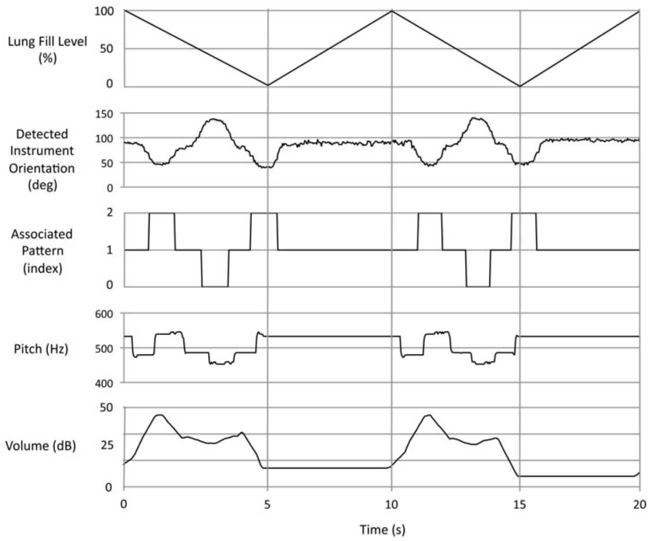

Fig. 4. In the advanced level interaction system's performance phase shown here, a professional musician controls the robot's output tone by changing the orientation of his instrument. In the graph the detected instrument orientation, the associated musical pattern and the output of the robot are displayed.

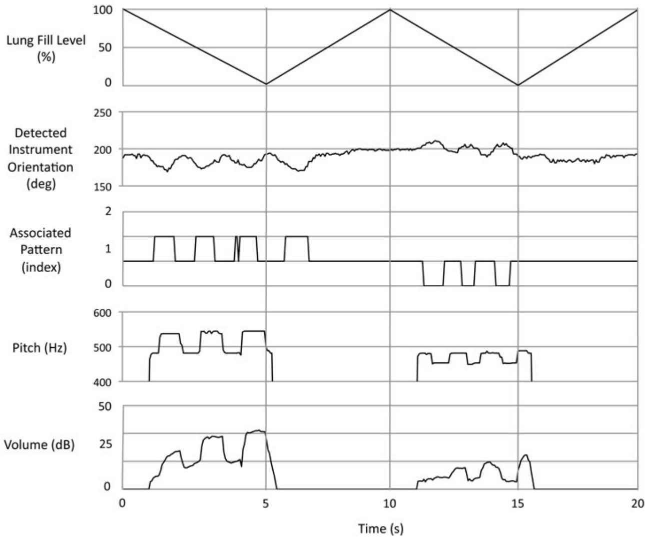

Fig. 5. Performance data for an intermediate level user. The user performs relatively fast movements, alternating between two notes for each breathing cycle.

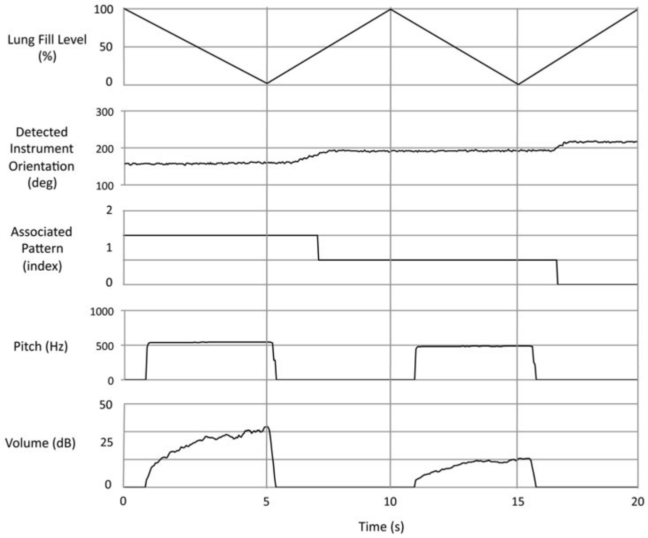

Fig. 6. The advanced interaction level used by a beginner instrumentalist. The user selects only a single note per breathing cycle.

3 melody patterns have been recorded.

The experiment was again performed by the professional flute player. After introducing the functionality of the system to the player, he performed one teaching phase and the following performance phase. In the following we show and evaluate an excerpt of the data recorded from the interaction of the professional player with the system.

The results for the advanced level interaction experiment in Fig. 4 show the output of the performance phase of the system. In the teaching phase the musician related three single notes $a4$, $b4$ and $c5$, to angles of 110°, 93° and 60°. The robot switches from the teaching phase to the performance phase after three notes / note patterns have been recorded. In the performance phase the musician varies the inclination angle of the flute (maximum: 146°, minimum 49°). With the inclination changing, also the note played by the robot changes as can be seen in the pitch analysis result plot (at $0.5s$, $2s$, $3s$, $4s$, $10.5s$, $10.5s$, $11.5s$, $12s$, $13s$, $14s$).

The same experiment was also performed with an intermediate and a beginner level player. The results are shown in Fig. 5 and Fig. 6. The beginner level does not fully use the capabilities of the interface by performing very slow movements,selecting only a single note for each breathing phase of the robot. The intermediate level player makes faster movements, but varies only between two notes for each cycle. Please note, that each of the players associated different instrument states to output notes during the respective teaching phases. To provide qualitative results documenting the usability of the system we performed the described experiments with two beginner-level, two intermediate-level and two professional level instrument players. We investigated their impression of the interaction quality with a questionnaire. This questionnaire asked the experiment subject to evaluate the system in three categories on a scale from 1 (= insufficient) to 10 (= excellent). The three categories to be questioned were evaluation of the Overall Responsiveness of the System, Adaptability to Own Skill-Level and Musical Applicability / Creative Inspiration. The result of the survey is shown in Fig. 7.

In the first category we questioned the Overall Responsiveness of the system to find out how the subjects responded to the technical implementation of the system in terms of detection and processing speed. We observed higher grades for the less experienced players and lower grades for the experienced players. With higher skill level the requirement for responsiveness seem to increase. The second evaluation category, Adaptability to Own Skill Level, was proposed to find out in how far the separation of the system in beginner level and advanced level interaction system fits for the

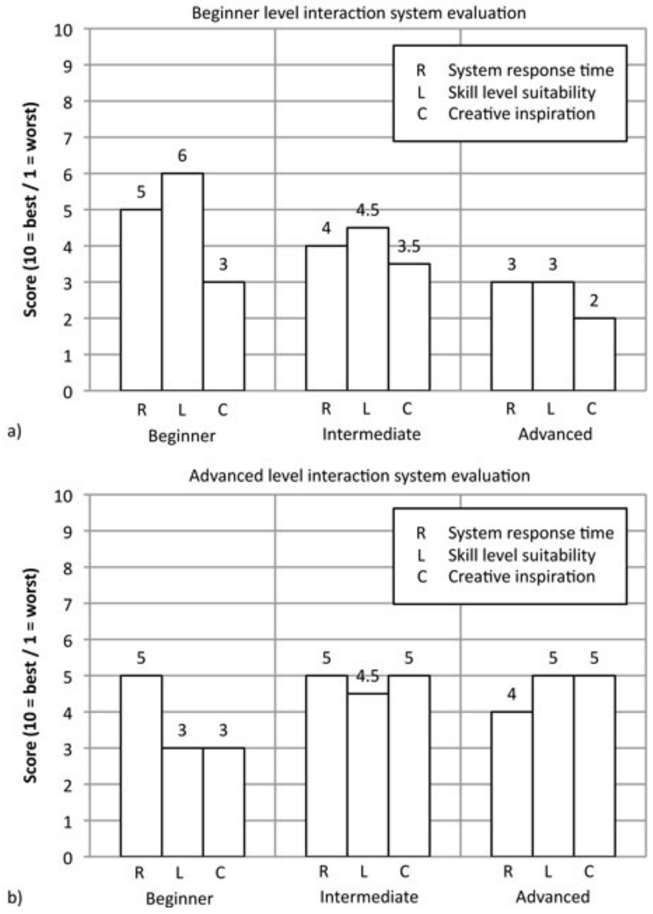

Fig. 7. Qualitative evaluation results for a) the beginner level interaction system and b) the advanced level interaction system. The results for each category were averaged over the number of subjects (two for each player skill level).

differently experienced players. We observed that according to our expectations the less experienced players would feel more comfortable with the beginner level interaction system and the more experienced players would give higher grades in case of the advanced level interaction system. In the Musical Applicability / Creative Inspiration section we tried to enquire about how the musicians felt they could express their musical intentions through utilizing the interaction interface. The results show intermediate scores for all skill levels.

The number of subjects used for the evaluation of the interaction system is with only 6 subjects very small. The experiments performed so far are only preliminary. We plan do perform experiments with more subjects as soon as possible.

IV. CONCLUSION AND FUTURE WORKS

The improved implementation of our two-stage mapping system makes it possible to interact with the robot in a more diverse way, compared to our previous mapping approach. Beginner level musicians can engage in interplay without having to consider about the physical restrictions of the humanoid. The state dependent mapping makes the robot aware of its own limitations and able to adjust its performance accordingly, similar to how a human player might act. We show in our experiments that this principle is applicable to simple improvisational play together with a musician partner. So far the possible modifications of the performance are limited to the fade-out produced by the low-pass filter module. In future works it might make sense, in order to finish a phrase before a lung-refill, to modify the currently played rhythm or melody. To do this, strategies on how a given pattern can be altered without the violation of any musical rules will be considered. Regarding the advanced level mapping module, we saw in the experimental results that an interacting musician can assign instrument gestures to musical expression variations.

Furthermore, we consider to try to make the system recognize the skill level of a performer automatically. The skill level could be assessed during an evaluation phase before the interaction, in which the player plays a musical phrase to the robot on his instrument. The robot after analyzing this data could automatically switch to the appropriate interaction level. Switching of interaction level during a performance might also be possible, however the evaluation of the interaction partner's skill level may be more significantly more difficult to implement.

REFERENCES

[1] J. Solis, K. Taniguchi, T. Ninomiya, and A. Takanishi, “Understanding the mechanisms of the human motor control by imitating flute playing with the waseda flutist robot wf-4riv,” *Mechanism and Machine Theory (Special Issue on Bio-Inspired Mechanism Engineering), Vol. 44 (3)*, pp. 527–540, 2008.

[2] J. Solis, K. Chida, K. Suefuji, and A. Takanishi, “The development of the anthropomorphic flutist robot at waseda university,” *International Journal of Humanoid Robots (IJHR)*, vol. 3, pp. 127–151, 2006.

[3] J. Fordham, “Jazz,” *DK Publishing (Dorling Kindersley), 1993*, 1993.

[4] G. Weinberg and S. Driscoll, “Towards robotic musicianship,” *Computer Music Journal, Vol. 30*, pp. 28–45, 2006.

[5] K. Petersen, J. Solis, and A. Takanishi, “Development of a aural real-time rhythmical and harmonic tracking to enable the musical interaction with the waseda flutist robot,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2303–2308, 2009.

[6] K. Petersen, J. Solis, and A. Takanishi, “Toward enabling a natural interaction between human musicians and musical performance robots: Implementation of a real-time gestural interface,” *The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 340–345, 2008.

[7] K. Petersen, J. Solis, and A. Takanishi, “Development of a real-time instrument tracking system for enabling the musical interaction with the waseda flutist robot,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 313–318, 2008.

[8] K. Sugiura and N. Iwahashi, “Learning object-manipulation verbs for human-robot communication,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2234–2240, 2008.

[9] S. Calinon and A. Billard, “A probabilistic programming by demonstration framework handling constraints in joint space and task space,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 367–372, 2008.

[10] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear / non-gaussian bayesian tracking,” *IEEE Transactions on Signal Processing*, pp. 174–188, 2002.