EXPRESSIVE HUMANOID ROBOT FOR AUTOMATIC ACCOMPANIMENT

Guangyu Xia[1], Mao Kawai[2], Kei Matsuki[2], Mutian Fu[1], Sarah Cosentino[2],
Gabriele Trovato[2], Roger Dannenberg[1], Salvatore Sessa[2], Atsuo Takanishi[2]

[1] Carnegie Mellon University, [2] Waseda University
{gxia, mutianf, rbd}@andrew.cmu.edu[1]
contact@takanishi.mech.waseda.ac.jp[2]

ABSTRACT

We present a music-robotic system capable of performing an accompaniment for a musician and reacting to human performance with gestural and facial expression in real time. This work can be seen as a marriage between social robotics and computer accompaniment systems in order to create more musical, interactive, and engaging performances between humans and machines. We also conduct subjective evaluations on audiences to validate the joint effects of robot expression and automatic accompaniment. Our results show that robot embodiment and expression improve the subjective ratings on automatic accompaniment significantly. Counterintuitively, such improvement does not exist when the machine is performing a fixed sequence and the human musician simply follows the machine. As far as we know, this is the first interactive music performance between a human musician and a humanoid music robot with systematic subjective evaluation.

1. INTRODUCTION

In order to create more musical, interactive, and engaging performances between humans and machines, we contribute the first automatic accompaniment system that reacts to human performance with humanoid robot expression (as shown in Figure 1). This study bridges two existing fields: *social robotics* and *automatic accompaniment*.

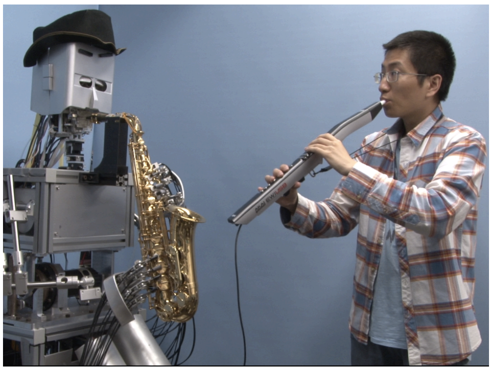


Figure 1. The robotic automatic accompaniment system.

On one hand, score following and automatic accompaniment systems (often briefly named automatic accompaniment) have been developed over the past 30 years to serve as virtual musicians capable of performing music with humans. Given a performance reference (usually a score representation), these systems take human performance as an input, match the input to the reference, and output the accompaniment by adjusting its tempo in real time. The first systems invented in 1984 [1][2] used simple models to anticipate the tempo of a monophonic input. Ever since then, many studies extended the model to achieve more expressive music interactions. These extensions include polyphonic [3] and embellished [4] input recognition, smooth tempo adjustment [5][6], and even expressive reaction with music nuance [7]. While most efforts focused on the system's auditory aspects, two major issues of automatic accompaniment remain unexplored. First, no model has considered the virtual musician's gestural and facial expressions, despite the fact that visual cues also serve as an important part of music interaction [8][9]. Second, no subjective evaluation has been conducted to validate that automatic accompaniment is a better solution than fixed media for human-computer music performance.

On the other hand, social robots have been developed to interact with humans or other agents following certain rules of social behaviors. Many studies have shown that robot expression, especially humanoid expression, significantly increases the engagement and interaction between humans and computer programs in many forms, such as telecommunication [10] and dialog systems [11]. However, music interaction, as high-level social communication, has not been paid much attention in this context. Though we have seen the development of several music robots, none are able to react to other musicians with human-like expression yet.

It is clear to see that automatic accompaniment and social robotics can complement each other. Therefore, we integrated the saxophonist robot developed at Waseda University into an existing framework of automatic accompaniment. To be specific, the system currently takes a human musician's MIDI flute performance as input and outputs acoustic accompaniment with gestural and facial expression. The (larger scale) gestural expression reacts to music phrases while the (smaller scale) facial expression reacts to local tempo changes. Of course, our first integration does not consider *all* aspects of gestural and facial expression. The current solution considers body and eyebrow movements, and we believe that other aspects of expression can be processed in a similar way.

In addition, we conducted subjective evaluations of this integration on audiences to validate the joint effects of robot expression and automatic accompaniment. Our hypothesis is that with humanoid robot expression, an automatic accompaniment system provides more musical, interactive, and engaging performance between humans and machines. We showed video clips in different conditions (with/without expression, with/without accompaniment) to audiences and used repeated-measure ANOVA to measure the difference between different conditions. Our results show that robot embodiment, especially facial expression, improves the subjective ratings on automatic accompaniment significantly. Counterintuitively, such improvement does not exist when the machine is performing a fixed media and the human musician simply follows the machine.

The next section presents related work. Section 3 presents the design of our saxophone robot with a focus on its control of body and eyebrow movements. Section 4 shows the automatic accompaniment framework with a focus on the mapping from MIDI performance to robot motions. In Section 5, we present the subjective evaluations and the experimental results.

2. RELATED WORK

Related work in music robotics can be categorized according to two perspectives: non-humanoid vs. humanoid, and pre-programmed vs. and interactive. Our study considers interactive humanoid robot.

2.1 Non-humanoid vs. Humanoid Music Robots

Musical player robots play an important role in the study of musical interaction. Non-humanoid music player robots with advanced interaction capabilities, such as Shimon [12] and Haile [13], have been used extensively to test pure musical interaction models. On the other hand, humanoid robots can be used as a tool for the validation of *embodied* interaction models. We believe that non-verbal gestures can be mimicked exquisitely and replicated by robots to study the influence of embodiment in musical interaction.

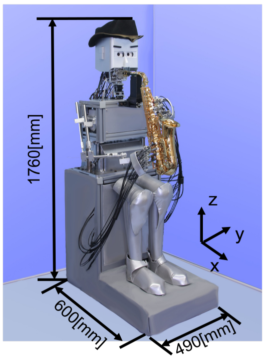
2.2 Pre-programmed vs. Interactive Music Robots

While most music robots have the potential to adapt their performance to others, most of their performances are still pre-programmed. However, we started to see more interactive music robots developed in the past decade. Generally, these robots detect beats from music and adjust their behaviors to stay synchronized with the music. These systems include interactive dancers [14], a Theremin player [15], singers [16], drummers [17], marimba players [12], and other percussion players [18]. However, very few of them react to music with gestural expression or have been evaluated experimentally by human subjects. As far as we know, the only subjective evaluation for interactive music robots was done in the work on Shimon [12]. This work showed that the visual contact with the marimba robot improves audiences' subjective ratings. On top of this, our study incorporates humanoid gestural and facial expression and conducts the first evaluation to inspect the joint effect of robot expression and music interaction.

3. HUMANOID SAXOPHONIST ROBOT

3.1 WAseda Saxophonist Robot (WAS)

The development of WAseda Saxophonist Robot (WAS) [19][20] was started in 2008. The robot was designed with a critical focus on the physiology and anatomy of the human organs involved during saxophone playing. The fourth version of the robot (WAS-4) was completed in 2015. Face and trunk mobility has been increased, to add basic interaction abilities during artistic joint performances with human partners. During joint musical performances, in fact, musicians cannot use vocal signs and must rely on non-verbal body communication for synchronization. The robot is now able to perform human-like non-verbal signaling, giving partner human players real-time cues on its interpretation, allowing for a better control over synchronization and improving the interaction experience as well as the overall joint musical performance. Figure 2 shows the general design of the robot used in this study.



Function	Body Parts	DoFs
Sound production	Lips	2
	Oral cavity	1
	Tongue	1
	Lung	Pump 1 Valve 1
Key stroke	Fingers	Left 8 Right 11
Body movement	Hip	1
Facial expression	Eyebrow	1
Total		29

Figure 2. The design of Waseda saxophonist robot (WAS).

3.2 Body and Eyebrow Movements

The two interactive movements used in this study are: swinging the upper body and raising/frowning eyebrows. The body positions during a swing movement are shown in Figure 3, where the robot starts from a neutral position (left), swings forward and backward (middle two snapshots), and finally comes back to the neutral position again (right). Eyebrow movements are illustrated in Figure 4, where the left one is neutral, the middle one is raised, and the right one is frowning.

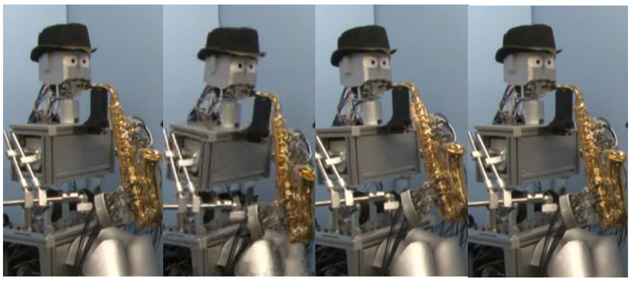


Figure 3. An illustration of the four positions in a body movement.

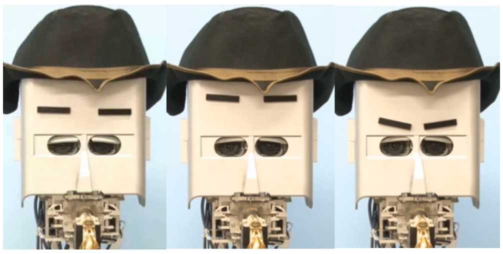


Figure 4. An illustration of the three eyebrow positions.

4. AUTOMATIC ACCOMPANIMENT WITH ROBOT EXPRESSION

This section describes how the robot reacts to human performance. There are three main steps: score following, tempo estimation, and the mapping from tempo to robot expression. The logic flow is shown in Figure 5. Again, the current system takes a human's monophonic MIDI flute performance as input and outputs acoustic accompaniment with eyebrow and body swing movements.

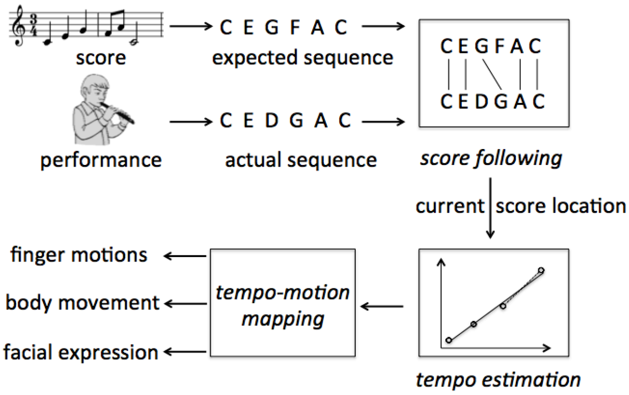


Figure 5. A system diagram of automatic accompaniment system with robot expression.

4.1 Score Following

Given the human performance, the first step of the process is score following, which keeps track of the current score location by finding the best match between score and performance. In our case, both score and performance are represented by a sequence of pitch symbols, with performance being the actual sequence and the score being the expected sequence. If the performance exactly follows the order of the score, we would simply update score location stepwise. However, since human performance will add and skip notes, we need an online matching algorithm. The current system adopts the solution introduced in [1], which first computes the "matched length" associated with each performance note and then updates the score location *only* when this length exceeds previous reported ones. Formally, the matched length is computed by:

$$\mathit{MatchedLength} = \#\text{ matched note} - \#\text{ skipped note} \quad (1)$$

Here, # represents the number of elements. Figure 6 shows an example, where the first line is the score, second line is the performance, and the third line is the matched length associated with each performed note. In this example, the algorithm reports a match and updates the score location after C, E, G, A, C are performed, since their matched lengths exceed previous ones.

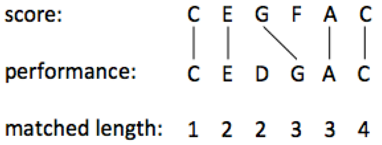


Figure 6. An illustration of the score following algorithm with one added and one skipped notes.

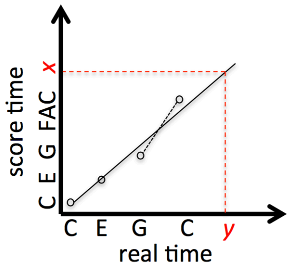


Figure 7. An example of the tempo estimation algorithm.

4.2 Temp Estimation

Given the matching results of score following, tempo estimation quantifies how fast/slow the human performance is against the timings specified in the score. This result will be later used to control the robot reactions. We adopt a "performance-score timing" 2-D representation (as shown in Figure 7) and represent tempi as slopes on this 2-D plane. The unit of score time is *beat*, the unit of performance time is *second,* and hence the unit of tempo is *beats per second*. We estimate tempo in two scales: micro and macro. The former is based on two adjacent notes, while the latter is based on the notes within a 4-beat interval.

Formally, let the matched notes reported by score following be $m = [m_1, m_2,\ldots, m_i,\ldots]$. Also, let the corresponding performance time and score time be $p = [p_1, p_2,\ldots, p_i,\ldots]$ and $s = [s_1, s_2,\ldots, s_i,\ldots]$, respectively. Then, the micro-scale tempo is defined as $v = [v_1, v_2,\ldots, v_i,\ldots]$, where

$$v_i = \begin{cases} (s_i - s_{i-1})/(p_i - p_{i-1}), & i > 1 \\ 1, & i = 1 \end{cases} \quad (2)$$

The macro-scale tempo is defined as $V = [V_1, V_2,\ldots, V_i,\ldots]$. If there are n matched notes within the score time interval of $[s_i - 4, s_i]$, then V_i is computed via the method of least squares:

$$V_i = \begin{cases} \frac{\sum_{j=i-n+1}^{i}(p_j-\bar{p})(s_j-\bar{s})}{\sum_{j=i-n+1}^{i}(p_j-\bar{p})}, & n > 1 \\ 1, & n = 1 \end{cases} \quad (3)$$

Here,

$$\bar{p} = \tfrac{1}{n}\textstyle\sum_{j=i-n+1}^{i} p_j \text{ and } \bar{s} = \tfrac{1}{n}\textstyle\sum_{j=i-n+1}^{i} s_j \quad (4)$$

Figure 7 shows an example of tempo estimation corresponding to the score following example in Figure 6, where the solid line represent the macro-scale tempo of the last matched note C, and the dotted line represents its micro-scale tempo. (Note that we do not estimate the tempi of unmatched notes.)

4.3 Mapping from Tempo to Robot Motions

We designed rule-based methods to control the robot motion by the estimated tempo. The current system separates robot motions into three groups: finger motions which are controlled by macro-scale tempo, body movements which are controlled by the deviation of macro-scale tempo, and eyebrow movements which are controlled by the deviation of micro-scale tempo. These rules are designed according to domain knowledge of music performance.

4.3.1 Finger Motions

Finger motions control the accompaniment, whose timings are specified in another pre-defined score that is synchronized with the score for human performance. The robot uses the latest macro-scale tempo estimation and extrapolates this tempo (slope) to estimate and schedule the next note. Figure 7 shows an example, where the nearest accompaniment note after the last human performed note C is at beat *x*, and its actual performance time will be *y*. It is important to notice that finger motions requires high timing accuracy, but robot mechanics has unavoidable latency. To overcome the latency, we schedule the notes ahead of their estimated onset times. Formally, if the latency for the MIDI flute played by the human performer is l_1 and the latency for the robot fingers is l_2, notes whose estimated time is t will be scheduled to execute at $t' = t - (l_2 - l_1)$. In practice, t' is around 40 milliseconds. (We point readers to [6] for more details on adjusting latency in real time performance.)

4.3.2 Body Movements

Body movements are controlled by the deviation of the macro-scale tempo. If the two latest estimated macro-scale tempi both speed up/slow down beyond a certain threshold, a body movement is triggered. By referring to the notations in the last section, for $i > 1$, a body movement is triggered if:

$$\left|\frac{v_{i-j} - v_{i-j-1}}{v_{i-j}}\right| > p \text{ for } j = 0 \text{ and } 1 \tag{5}$$

The rationale of this rule is that performers often use body movements to indicate smooth tempo changes. The current system sets $p = 5\%$. Besides this rule, we also insert a body movement at the beginning and the ending of the robot performance.

4.3.3 Eyebrow Motions

Eyebrow motions are controlled by the deviation of the micro-scale tempo. If the two latest estimated micro-scale tempi both slow up beyond a certain threshold, both eyebrows will raise. Similarly, if the tempi speed up beyond a certain threshold, a *frown* motion is triggered. If none of these two conditions are met, eyebrows stay at the neutral position. Formally, for $i > 1$, and $j = 0$ and 1,

$$\text{eyebrow motion} = \begin{cases} \text{raise}, & \text{if } \frac{v_{i-j} - v_{i-j-1}}{v_{i-j}} > q \\ \text{frown}, & \text{if } \frac{v_{i-j} - v_{i-j-1}}{v_{i-j}} < -q \\ \text{neutral}, & \text{otherwise} \end{cases} \tag{6}$$

The rationale of this rule is that eyebrow motions are often associated with sudden tempo changes. The current system sets $q = 5\%$. Note that eyebrow motions will be more frequent than body movements under the same threshold because micro-scale tempi are more sensitive.

5. SUBJECTIVE EVALUATION

We conducted subjective evaluations on audiences to validate the effects of robot expression. We first inspected whether the robot helps with automatic accompaniment. Then, we inspected whether the robot helps with fixed media performance (in which the robot plays a fixed performance and the human performer has to adapt to the robot). Finally, we compared these two results to see the joint effect of robot expression and automatic accompaniment.

5.1 The Robot Effect on Automatic Accompaniment

Our hypothesis is that with humanoid robot expression, the automatic accompaniment system provides more musical, interactive, and engaging performance between humans and machines. To test this claim, we recorded videos of human-computer interactive performances (of the same piece of music) in 3 different conditions of robot embodiment and invited audiences to provide subjective ratings on these videos.

5.1.1 Video Recording Setup

The videos were recorded as shown in Figure 8, with the human performer and the robot standing opposite to each other. Figure 1 shows a corresponding screen shot.

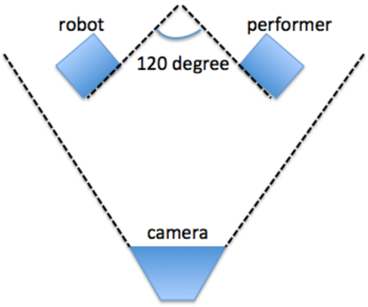


Figure 8. The layout of the video recording setup.

Index	Robot setting
A	Blocked robot
B	Static body
C	Full expression

Table 1. The three different conditions for robot setting.

5.1.2 Conditions of Robot Embodiment

The 3 performance conditions are listed in Table 1, where higher index corresponds to greater functionality of the robot. Note that in condition A (blocked robot), we put a cover in front of the robot so that neither the human performer nor the camera could see the robot. The purpose was to block the visual cues but retain the same sound source. In condition B (static body), the robot's body and eyebrows do not move; the only working parts are the mouth and fingers. In condition C (full expression), the robot movements include mouth, fingers, body, and eyebrows.

5.1.3 Survey

We showed the recorded performance videos in all 3 conditions in a random order to each audience subject without directly revealing the condition. Each video is about 80 seconds long. After each video, audiences were asked to rate the performance according to three criteria:

1. *Musicality:* how musical the performance was.
2. *Interactivity:* how close the interaction was between the human performer and the machine.
3. *Engagement:* how engaged the human performer was.

For all 3 criteria, we used a 5-point Likert scale from 1 (very low) to 5 (very high).

5.1.4 Hypothesis Test

The null hypothesis is that different conditions have no effect on automatic accompaniment and therefore the ratings under different conditions are the same. Formally:

$$H_0: \quad \mu_A = u_B = \mu_C \tag{7}$$

Similarly, the alternative hypothesis is that:

$$H_1: \quad \exists i,j \in \{A,B,C\}: \mu_i \neq \mu_j \tag{8}$$

Since all the subjects experienced all the conditions, we used within-subject ANOVA [21] (also known as repeated measurement study) to compute the mean standard error (MSE) and p-value. We used the Huynh-Feldt correction [21] when the sphericity of the data is not met.

5.1.5 Results

A total of *n* = 33 subjects (14 female and 19 male) have completed the survey. The aggregated result (as in Figure 9) shows that the robot effect improves the subjective ratings of automatic accompaniment.

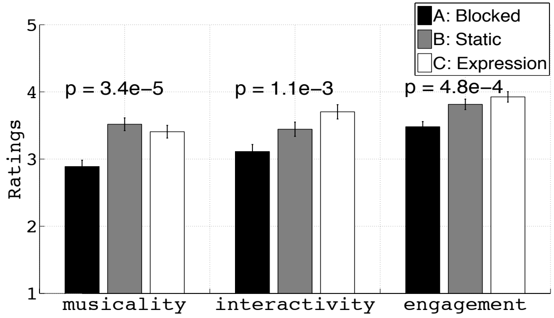


Figure 9. The subjective evaluation results of the robot effect on automatic accompaniment.

Here, different colors represent different conditions. The heights of the bars represent the means of the ratings and the error bars represent the MSEs. It is clear that the robot embodiment and expression improves the ratings and such improvements are monotonic (except for *musicality*) when the functionality of the robot increases. For all three criteria, the p-values are much smaller than 0.005 and hence the improvements are statistically significant.

5.2 The Robot Effect on Fixed Media Performance

In addition to the robot effect on automatic accompaniment, we also inspected whether the robot helps with fixed media performance. In this case, the robot played a pre-recorded performance and the human musician adapted to the robot. Similar to Section 5.1, the null hypothesis is that different conditions have no effect on the subjective ratings of fixed media performance. With exactly the same video recording setup, conditions of robot setting, and survey process, the result (as in Figure 10) shows that robot embodiment and expression do not help with fixed media performance.

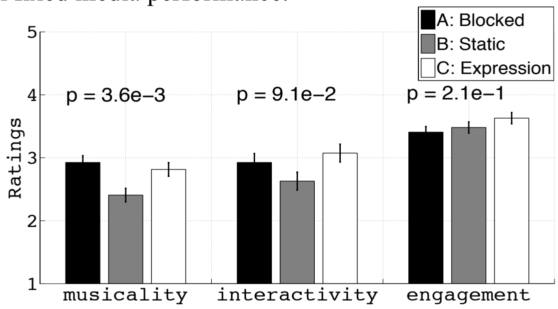


Figure 10. The subjective evaluation results of the robot effect on fixed media performance.

Counterintuitively, for musicality the robot decreases the ratings with the p-value smaller than 0.005. For interactivity and engagement, though we see evidence of improvement, the associated p-values are both larger than 0.05.

5.3 A Comparison between Automatic Accompaniment and Fixed Media Performance.

We finally inspect the joint effect of automatic accompaniment and the robot effect by putting the results of Figure 9 and Figure 10 together, as shown in Figure 11.

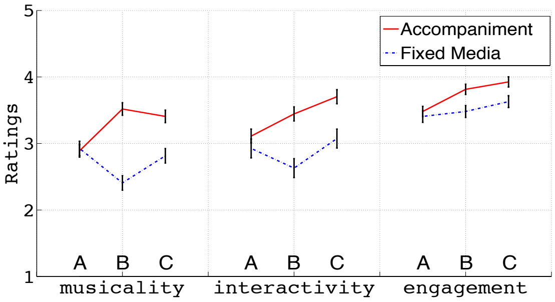


Figure 11. The joint effect of automatic accompaniment and robot expression.

For all 3 criteria, while the difference between automatic accompaniment and fixed media is not significant (p-value > 0.05) in condition A, the difference becomes much more significant (p-value < 0.005) in conditions B and C. This result suggests that neither automatic accompaniment nor robot expression alone is significantly better than fixed media performance. Only when we combine these two factors, does the music performance between the human and the machine become significantly more musical, interactive, and engaging.

6. CONCLUSIONS AND FUTURE WORK

In conclusion, we have combined the efforts of social robotics and automatic accompaniment to create the first automatic accompaniment system with humanoid robot expression. Our subjective evaluation shows that expressive humanoid robots lead to more musical, interactive, and engaging automatic accompaniment. Counterintuitively, this effect does not exist for fixed media performance. This study contributes to the computer music community by providing the first subjective evaluation on automatic accompaniment and its joint effect with robot expression. It also contributes to the social robotics community by proving that the benefits of humanoid robots generalize to the interactive music performance scenario. The result shows the benefit of combining interactive computer music system with humanoid robot, which points to the integration of these two fields for future research.

In the future, we would like to continue this study in the following three directions:

Visual cues: The current robot is still blind. We are going to place cameras on the robot and use visual cues to guide the generation of robot expression.

Learning-based robot expression: So far, the robot expression is generated by a rule-based method. To make the method learning-based, we are going to use a motion capture system to collect rehearsal videos with facial and gestural expression.

Evaluation of the performance experience: So far, the subjective evaluation is conducted on audiences only. We are going to invite multiple performers as our subjects in the future.

Acknowledgments: We thank to the General Directorate for Cultural Promotion and Cooperation for its support to RoboCasa, and we also thank to SolidWorks Corp for their support to the research.

7. REFERENCES.

[1] R. Dannenberg, “An online algorithm for real-time accompaniment,” in Proceedings of the International Computer Music Conference, 1984, pp. 193-198.

[2] B. Vercoe, “The synthetic performer in the context of live performance,” in Proceedings of the International Computer Music Conference, 1984, pp. 199-200.

[3] J. Bloch and R. Dannenberg, “Real-time accompanyment of polyphonic keyboard performance,” in Proceedings of the International Computer Music Conference, 1985, pp. 279-290.

[4] R. Dannenberg, and H. Mukaino, “New techniques for enhanced quality of computer accompaniment,” in Proceedings of the International Computer Music Conference, 1988, pp. 243–249.

[5] A. Cont., “ANTESCOFO Anticipatory Synchronization and Control of Interactive Parameters,” In Computer Music Proceedings of International Computer Music Conference, 2008.

[6] D. Liang, G. Xia, and R. Dannenberg, “A framework for coordination and synchronization of media,” in Proceedings of the New Interfaces for Musical Expression, 2011.

[7] G. Xia and R. Dannenberg, “Duet Interaction: learning musicianship for automatic accompaniment,” in Proceedings of the International Conference on New Interfaces for Musical Expression, 2015.

[8] K. Katahira et al., “The role of body movement in co-performers’ temprral coordination,” in Proceedings of ICoMCS, 2007, pp. 72.

[9] S. Kawase, “Importance of communication cues in music performance according to performers and audience,” International Journal of Psychological Studies, 2014, pp. 49-64.

[10] S. Adalgeirsson, “Mebot, a robotic platform for socially embodied telepresence,” in Proceedings of the fifth ACM/IEEE International Conference on Human-Robot Interaction (HRI-10), ACM/IEEE International, 2010, pp. 15-22.

[11] J.K. Lee and C. Breazeal, “Human social response toward humanoid robot’s head and facial features,” CHI Extended Abstracts, 2010, pp. 4237-4242

[12] G. Hoffman, and G. Weinberg, "Interactive improvisation with a robotic marimba player," in Autonomous Robots 31, no. 2-3, 2011, pp. 133-153.

[13] S. Sun, M. Trishul, and G. Weinberg. "Effect of Visual Cues in Synchronization of Rhythmic Patterns." In Proceedings of International Conference of Music Perception and Cognition, 2012.

[14] G. Xia, et al., "Autonomous robot dancing driven by beats and emotions of music," In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. 2012, pp. 205-212.

[15] A. Lim, et al., "Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist," IEEE International Conference on Intelligent Robots and Systems, 2010, pp. 1964-1969.

[16] T. Otsuka, et al., "Incremental polyphonic audio to score alignment using beat tracking for singer robots," IEEE/RSJ International Conference on. IEEE, 2009, pp. 2289-2296.

[17] G. Weinberg, S. Driscoll, and M. Parry, “Musical interactions with a perceptual robotic percussionist,” in IEEE International Workshop On Robot and Human Interactive Communication, 2005, pp. 456-461.

[18] A. Kapur, “Multimodal techniques for human/robot Interaction.” In Musical Robots and Interactive Multimodal Systems, Springer Berlin Heidelberg, 2011, pp. 215-232.

[19] J. Solis, T. Ninomiya, K. Petersen, M. Takeuchi, and A. Takanishi, “Development of the anthropomorphic saxophonist robot WAS-1: Mechanical design of the simulated organs and implementation of air pressure feedback control”. In Advanced Robotics, 24(5-6), 2010.

[20] K. Matsuki, K. Yoshida, S. Sessa, S. Cosentino, K. S. Kamiyama, and A. Takanishi “Facial Expression Design for the Saxophone Player Robot WAS-4”, In proceedings of the 21st CISM IFToMM Symposium on Robot Design, Dynamics and Control, 2016.

[21] R. Ellen and E. Girden. “ANOVA: Repeated measures”. No. 84. Sage, 1992.