A musical robot that synchronizes with a co-player using non-verbal cues

Angelica Lim, Takeshi Mizumoto, Tetsuya Ogata, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan

{angelica, mizumoto, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

Music has long been used to strengthen bonds between humans. In our research, we develop musical co-player robots with the hope that music may improve human-robot symbiosis as well. In this paper, we underline the importance of non-verbal, visual communication for ensemble synchronization at the start, during and end of a piece. We propose three cues for inter-player communication, and present a theremin-playing, singing robot that can detect them and adapt its play to a human flutist. Experiments with two naive flutists suggest that the system can recognize naturally occurring flutist gestures without requiring specialized user training. In addition, we show how the use of audio-visual aggregation can allow a robot to adapt to tempo changes quickly.

keywords: entertainment robots, gesture recognition, audio-visual integration

1 INTRODUCTION

In Japan's aging society, the elderly may soon rely on robots to perform chores or assist in day-to-day tasks. According to one survey, one of the main requirements for these robot companions is natural, human-like communication [1]. Indeed, if a robot lacks communication skills, the human may sense a feeling of incompatibility, fear, and frustration [2]. Especially in cases where a task may involve a human's safety, a certain trust is needed between human and robot; this necessity is often referred to as *human-robot symbiosis*. Therefore, it's essential to find ways of building a bond of familiarity and trust between humans and robots if we want them to be accepted as helpers in our society.

Music has a long history of creating social bonds between humans. Every culture in the world has gathered from time to time in groups to participate in rhythmic song or dance. As stated by psychologist Brown [3], "music is an important device for creating group-level coordination and cooperation[:] a musical performance [...] tends to create a symbolic feeling of equality and unity, one that produces a leveling of status differences among the participants, thereby dampening within-group competition." Indeed, in a study by Wiltermuth et al. [4], it was shown that groups that sang together trusted each other more in a subsequent prisoner's dilemma game.

Musicologists say it is the *synchrony* of music and dancing that induces this social bonding effect [3]. Even in the era of proto-humans, some things could only be possible with synchronized movement: to transport heavy stones, people had to pull in synchrony; by shouting together, the sound could be projected farther away. McNeill [5] observed that the synchronized march of soldiers is still being practiced despite its practical uselessness in modern times; he reasoned that this "muscular bonding" is important for group cohesion. Recent studies in biology may support this observation; it was shown that rowers had a higher pain threshold when rowing in unison with other rowers, rather than in the individual condition [6]. The mirroring or "chameleon" effect may also be at play here: a subject is more likely to get along harmoniously with a group if he/she acts similarly. The reverse is true if the subject acts out of sync [7] [8].

Our goal is thus to improve human-robot symbiosis by making a musical co-player robot, focusing particularly on the synchronization aspect. We focus on a score-playing robot, for example to play in duets or trios [9] based on a written score. In ensembles, temporal coordination is crucial because expressive music contains many deviations from the score; for example, a player may speed up their play to brighten the musical mood, or slow down to express sadness [10]. It is during these changes when the robot must synchronize with the human player's tempo. In addition, we focus on two points in music where synchronization is key: the start of the first note, and the ending.

Let's briefly survey the systems which already exist for digitally accompanying musicians. In the field of computer accompaniment, play-back software such as [11] [12] track the notes played by the musician to determine where to play in its own score. This is known as score following, and has also been implemented by Otsuka et al. [13] on a robot system. However, it has been suggested [14] that this is not how humans keep in time; musicians cannot, and do not need to, memorize co-players' scores to stay in sync. A more likely explanation is that musicians keep track of the music's "pulse". This approach is called beat tracking. For example, Georgia Tech's HAILE drum robot [15] detects human drumbeats using energy-based beat trackers. Using the beats, it can detect speed and perform improvisation accordingly. In [16], a robot can listen to pop music and sing along to the beat. A common problem is that these beat trackers have difficulty when there is no percussive, steady beat, for instance in violin or flute ensembles. So how do humans synchronize in these situations?

In real ensembles, nonverbal behaviors like body movement, breathing, and gaze are used to coordinate between players [17]. Multiple studies show the importance of visual information. In [18], Wii-mote-carrying children danced to music, and were shown to move with better synchronization in a face-to-face social condition as opposed to dancing alone. Katahira et al. [19] compared pairs of face-to-face and non-face-to-face drummers, and also found a significant contribution of body movement to temporal coordination between performers. By observing authentic piano duets, [20] found that head movements, exaggerated finger lifts and eye contact are used to communicate synchronization events between players. Finally, Fredrickson [21] showed that band musicians synchronize best by *both* watching the conductor and listening to their co-players. Truly, both audio and vision are important for synchronization.

Figure 1: The singing, theremin-playing music robot detects a flutist's cues to play a duet.

In this paper, we describe a unique singing, theremin-playing robot that can synchronize using both these senses. We assume a small human-robot ensemble (e.g. duet or trio) with no conductor. In this case, we need to formalize the musician-to-musician communication between humans, a topic little studied in music literature so far. In particular, we posit that there exist inter-player *cues* for coordinating at least three types of temporal events: start, stop, and tempo change. We evaluate a music robot system that plays the theremin [22] and sings, playing music with a flutist in the following way:

(1) It begins playing when it detects a visual cue from the flutist

(2) It changes its tempo by watching and listening to the flutist

(3) It ends a held note (i.e. fermata) when visually indicated

This robot co-player system shown in Fig. 1 has been described in detail in [23] [24]. In this paper, we review the components of the system and examine its validity with naive users.

1.1 VISUAL CUES

We first formalize the concept of *visual cue*, hereafter also called *gesture*. In conducting, a typical gesture denotes the tempo at which the musicians should synchronize. These visual events have been called "beats" in [25], and are typically described as the conductor baton's change in direction from a downward to an upward motion [26]. Outside of traditional conducting studies, research on clarinetist's movements found that "movements related to structural characteristics of the piece (e.g. tempo)" were consistently found among player subjects, such as "tapping of one's foot or the moving of the bell up and down to keep rhythm" [27]. This up-and-down motion will be the basis of the visual cues described next.

According to our informal observation, flutists move their flutes up and down in a similar way to a conductor's baton to communicate within an ensemble. Our three observed cues are shown in Fig. 2.

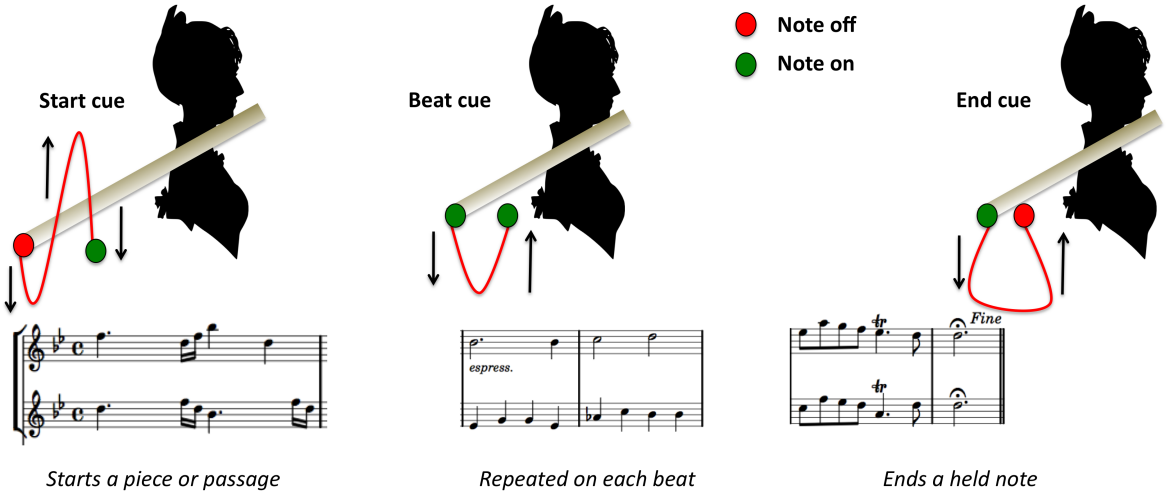


Figure 2: Trajectories of flute visual cues, along with examples of locations used in score. As shown, the end cue is a circular movement of the end of the flute, and the beat cue is a simple down-up movement. Despite this difference, they both appear as a down-up motion when viewed from the front.

A DOWN-UP-DOWN motion of the end of the flute indicates the start of a piece, while the bottom of a DOWN-UP motion, called an *ictus* in conducting, indicates a beat. Finally, a circular motion of the end of the flute indicates the end of a held note. We hypothesize that players of other baton-like instruments like clarinet, trumpet or trombone may also use similar signals to communicate. Here, we verify whether these cues are used naturally between flutists.

We define "natural" here as "without needing explicit prompting". This is in opposition with system-specific gestural control. Consider the flute-tracking computer accompaniment system in [28] which plays a given track when the flutist makes a pre-defined pose with her flute, for example "pointing the flute downward and playing a low B." This gesture is system-specific, and not a natural gesture used among real flute players. The advantage of detecting natural gestures is that the users do not have to learn nor think about special movements to control the robot, which can be difficult when already occupied with performing a piece. In addition, other human co-players will also "naturally" understand the flutist's cues, making the ensemble size scalable.

2 A ROBOT CO-PLAYER SYSTEM

In this section, we describe how to recognize the visual cues shown in Fig. 2. We will then describe the shortcomings of a purely visual system and how we augment it with audio. Finally, we will give an overview of our robot co-player system.

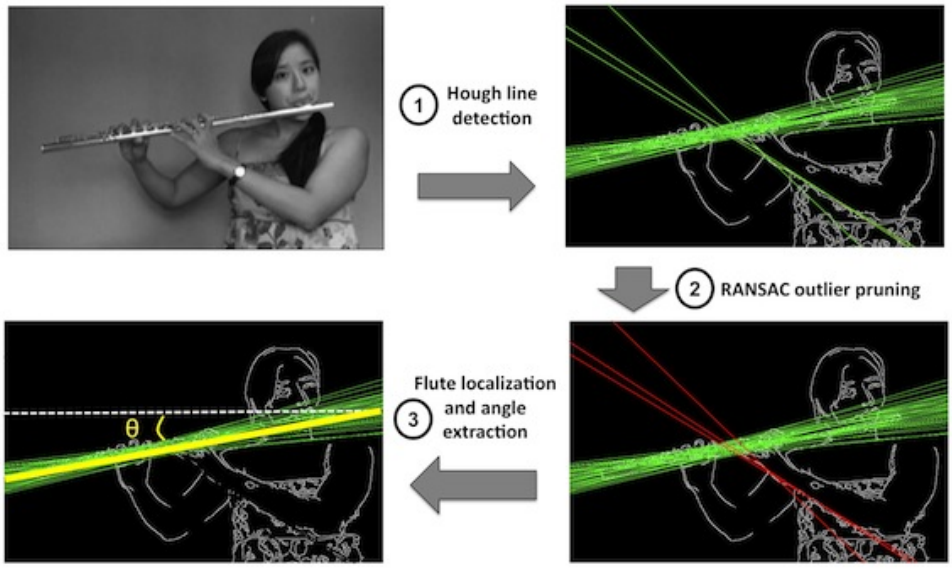


Figure 3: Original input image (top left), detected Hough lines (top right) and outliers marked in red (bottom right).

2.1 DETECTING VISUAL CUES

2.1.1 Flute localization

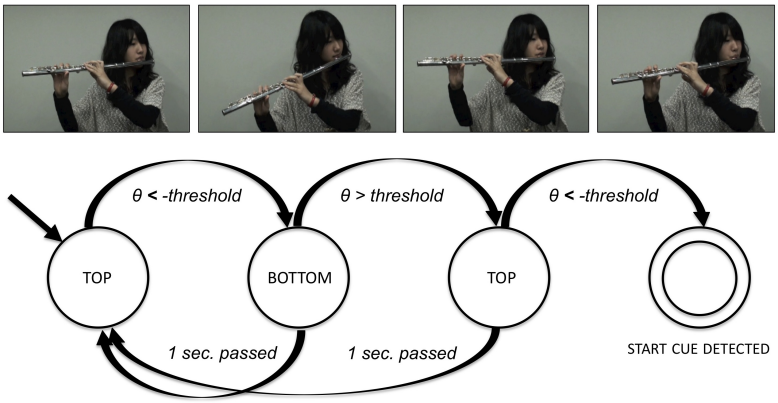
The first step in detecting the visual cues described in Sec. 1.1 is localizing the flute. The process is shown in Fig. 3. In our system, we assume the robot faces the flutist such that its camera produces images like Fig. 3 (top left). Localization is performed by using a combination of Canny edge detection [29], the Hough transform [30] and RANSAC outlier pruning [31]: the Hough line detection algorithm outputs many lines along the flute, and RANSAC removes spurious lines caused by background or clothing. The flute angle θ is calculated as the mean of the angles of the remaining inlier lines. Other tracking methods such as optical flow may be considered for a more generic system; we selected this simple angle-extraction approach to be robust against noise caused by camera movement while the robot plays the theremin.

2.1.2 Flute tracking

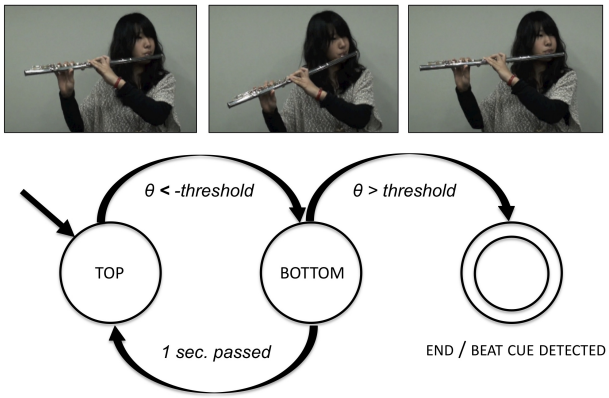
Next, our system tracks the flute angle calculated from the localization step. For each pair of consecutive video frames F at time $t-1$ and t, we calculate the change in θ:

$$\Delta\theta = \theta(F_t) - \theta(F_{t-1}). \tag{1}$$

The flute's speed, defined here as $\Delta\theta$, is input into the finite state machines (FSM) in Fig. 4. Notice that the beat cue and end cue FSMs are the same due to their similarity when viewing the flutist from the front. When the end of the flute is moving downwards faster than a certain threshold, the FSM moves into a BOTTOM state, and conversely it moves into a TOP state. The speed threshold acts as



(a)



(b)

Figure 4: Finite state machines for start cue (a) and end/beat cues (b).

a basic smoother, so that the state is not changed for small values of θ. We reset the FSM state to the beginning if no significant motion is detected for 1 second.

2.1.3 Using visual cues during a performance

Context is important for deciding what a movement means. For example, a hand wave could both be used to say goodbye, or to shoo away a fly. In our system, we filter our visual cues based on context; here, context is based on score location. Start cues only control the robot's play at the start of the piece, and end cues are only given attention when the robot is currently holding a note. Contrary to the start and end cues, beat cues are valid throughout the piece.

Beat cues are used to detect changes in tempo. Our initial tempo change mechanism [32] required the player to perform three regularly-spaced visual beat cues to indicate a tempo change. The average difference between these beat cues determined the tempo. This three-cue sequence ensured the movements were indeed purposeful messages to the robot to change tempo, and not arbitrary movements while playing. The drawback of this approach is that performing three regularly-spaced beat gestures is too strenuous for continued use throughout a performance. The method described in the following sections integrates audio cues such that only two beat gestures are required, as long as they are supported by audio information.

2.2 NOTE ONSET DETECTION

We use flute note onsets as our source of audio information. The term 'onset' refers to the beginning of a note, and onsets are useful for our system because notes may also indicate beats. For example, four consecutive quarter notes in a 4/4 piece would have a one-to-one correlation with the beats. Similarly, if there were more than four notes, the onsets could indicate a super set of the beats.

How can we detect note onsets? The review in [33] provides a good overview of methods, and selecting an appropriate note onset detector depends on our usage. For instance, our first requirement is that we want our robot to play in musical ensembles with a woodwind instrument – the flute. In this case, the note onset detection method must be more sensitive than those used for percussion instruments such as piano. It should detect soft tonal onsets; this includes (1) slurred changes between two pitches and (2) repeated notes of the same pitch. A conventional approach may include a detectors for each of these cases: a pitch detector for (1), and an energy-based [34] or Phase Deviation [35] detector for (2). We selected a method that can deal with both cases simultaneously by detecting changes in both spectral magnitude and phase in the complex domain [36].

Speed is also a requirement for our note onset detection method. We used the Aubio library [37] implementation of Complex Domain onset detection, which is written in C and calculates the Kullback Leibler divergence [38] in the complex domain from frame to frame in real-time. It should be noted, however, that as mentioned in [33], phase-tracking methods including this Complex Domain method are sensitive to noise, which we experienced when testing this on lower quality audio setups. Ideally, the

own robot's microphone should be used, with sound separation or frequency filtering should be used to separate the flutist's notes from the theremin sound (e.g. using the separation approach in [39]). For the present work, our combination of a lapel microphone and Complex Domain detection worked well, though a different method may needed if faced with audio signals containing environmental noise.

2.3 PERCEPTUAL MODULE: AUDIO & VISUAL BEAT MATCHING

The perceptual module of our co-player system combines the audio note onsets described in the previous section with the visual beat cues from Sec. 2.1. We assume that the flutist wants to change the tempo if: (1) the flutist plays notes on two consecutive beats, (2) makes visual beat cues on those beats, and (3) the beats indicate a tempo within pre-defined limits. This is consistent with how humans play - they do not, for example, triple their speed suddenly, unless it is already marked in the score.

We define instantaneous tempo as the time between the onset of the latest two beats, also known as Inter-Onset-Interval (IOI). Our algorithm for IOI extraction works as follows. Let V and A respectively be temporally ordered lists to which we add observed video and audio cue events at times t_v and t_a. When a given audio and visual cue are less than δ_1 milliseconds apart, we add the audio cue time to M, a temporally ordered list of *matched beat* times. We return a new tempo using the difference between the last two *matched beats*, as long as it differs no more than δ_2 milliseconds from IOI_c, the current tempo. Otherwise, we check whether the player has performed three beat cues resulting in two IOI's that differ by less than δ_3 (set to 1000ms in our experiments). If so, we return their average as the new IOI, under the same δ_2 tempo change constraint.

Whenever an audio or visual cue event at time e is detected at time t_e, we run the following function.

if e is *audio* **then**
  $A \leftarrow A + t_e$
  if $\exists v \in V, |t_e - t_v| < \delta_1$ **then**
    $M \leftarrow M + t_e$
    if $|M| \geq 2$ **and** $||M[last] - M[last-1]| - IOI_c| < \delta_2$ **then**
      return $M[last] - M[last-1]$
if e is *video* **then**
  $V \leftarrow V + t_e$
  if $\exists a \in A, |t_e - t_a| < \delta_1$ **then**
    $M \leftarrow M + min(\{t_a | a \in A, |t_e - t_a| < \delta_1\})$
    if $|M| \geq 2$ **and** $||M[last] - M[last-1]| - IOI_c| < \delta_2$ **then**
      return $M[last] - M[last-1]$
  if $|V| \geq 3$ **and** $(V[last] - V[last-1]) - (V[last-1] - V[last-2]) < \delta_3$ **then**
    if $(V[last] - V[last-2])/2 - IOI_c) < \delta_2$ **then**
      return $(V[last] - V[last-2])/2$

In short, visual beat cues can be viewed as an enable mask for the audio data. As shown in Fig. 5,

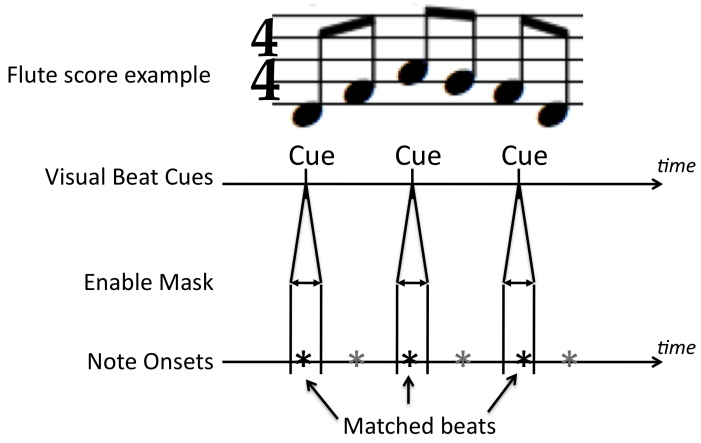


Figure 5: Our audio-visual matching scheme. Visual cues act as a filter for note onsets that fall into a given range around the visual cues. For a tempo change to be detected, only two of the three matched beats above are needed.

extraneous offbeat notes are filtered using with a window width of $2 * \delta_1$ around each visual beat cue. A *matched beat* corresponds to the note onset event that falls within that window. We experimentally set our threshold δ_1 to 150 ms, which gives a detection window of 300 ms around each visual beat cue. If more than one audio note onset is detected within this window, the first onset is chosen - the earliest onset detected.

It can be noted that the final IOI resulting from audio-visual matching is determined solely by the audio note onset time. This is due to audio signals' high sampling rate – we sample audio at 44100 kHz, whereas video camera outputs 30 frames per second. Thus, although audio data may contain unneeded note onsets (such as those at the bottom of Fig. 5), it is more precise. This precision is important, for example, when using more than one camera (e.g., with two robot co-players). Even minute differences in video frame rates and capture times can produce relatively large differences in detected tempos using a vision-only approach.

In order for this simple fusion algorithm to be valid, a precise timing scheme is essential. We chose to use Network Time Protocol [40] to synchronize the clocks of all our modules, some of which were connected through ethernet. Alternatively, the Carnegie Mellon laptop orchestra [41] used a central hub from which laptop instruments queried the current time. In addition to precise clock synchronization, this event-driven formulation of the algorithm is required because the data from two data sources may not arrive in sequence, due to network delays.

2.4 SYSTEM OVERVIEW

This system was implemented on the HRP-2 theremin-playing robot first introduced in [22], with the addition of a VOCALOID singing module [42]. Fig. 6 overviews the robot co-player system. The HRP-2's Point Grey Fly camera is used to take greyscale images at 1024x728 resolution, at a maximum

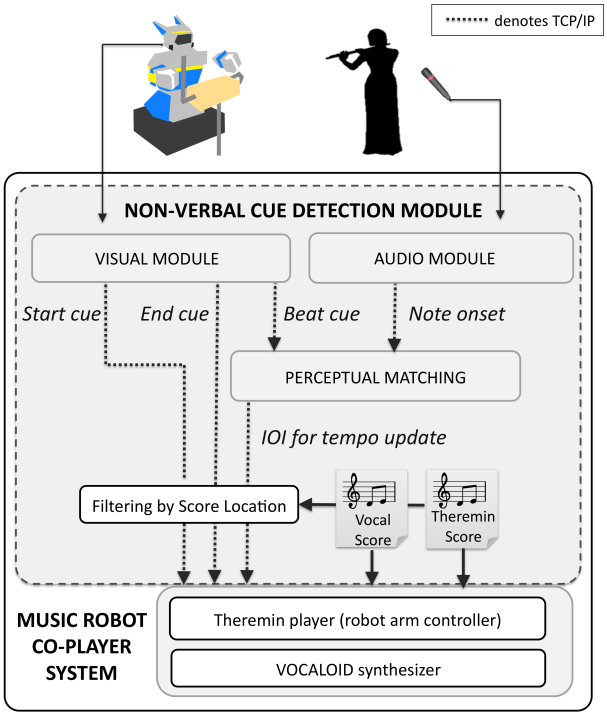


Figure 6: Overview of our robot co-player system

of 30 fps. When start and end cues are detected from the vision module, these commands are sent to the theremin robot to start a piece or end a held note, depending on the robot's current location in the score. A 2.13 GHz MacBook with an external microphone was used as our note onset detection module. The initial tempo is set to the one written in the score. After that, the system attempts to match input from its two input modalities within the perceptual matching module, and sends on detected tempos to the theremin player. As shown in Fig. 6, our non-verbal cue detection module controls two different music systems via the network: the theremin robot and a VOCALOID singing synthesis program. This suggests the portability of this system to other music tools with few or minor changes.

3 EXPERIMENTS AND RESULTS

We performed two experiments to determine the viability of our co-player system. Experiment 1 evaluates the start and stop gesture recognition module. Experiment 2a and 2b evaluate the tempo tracking functionality.

3.1 Experiment 1: Visual cue detection with naive players

In [23], we found that our method detected start cues at greater than 93% accuracy, and end cues with 99% accuracy given our initial study with one flutist. In this experiment, we recruited two naive flutists

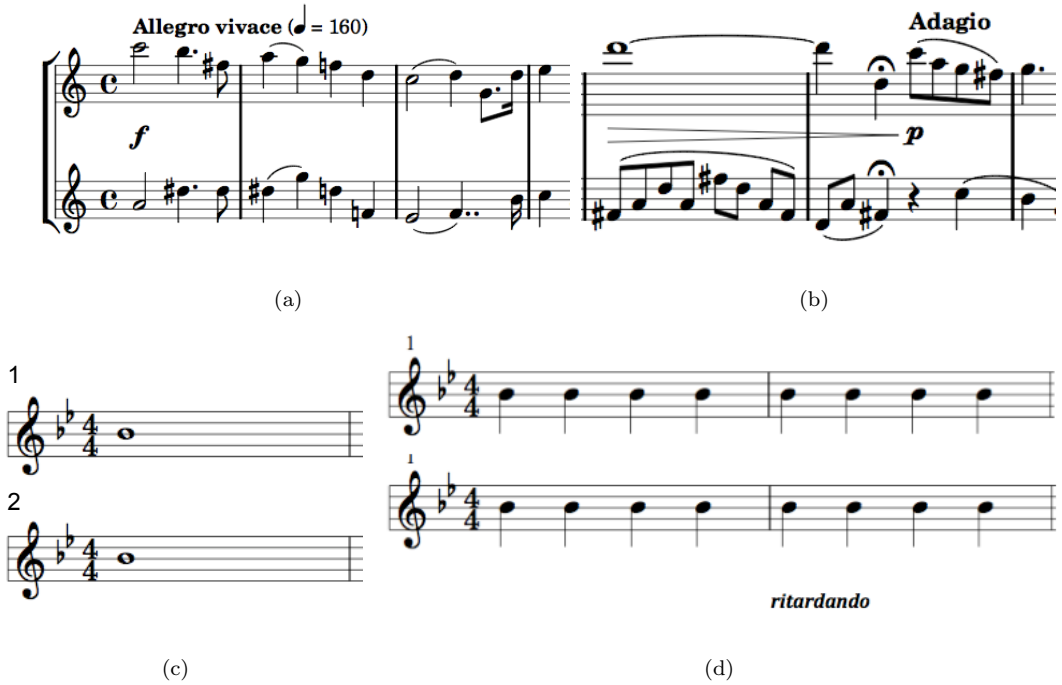


Figure 7: Musical situations used for surveys: a) the beginning of a duet to investigate start cue b) a passage with fermata for end cue c) a note with simultaneous start and stop and d) a ritardando passage to investigate beat cues.

from Kyoto University's music club to further evaluate our system. Participant A was an 19-year-old female with 12 years of flute-playing experience, and participant B was a 22-year-old female player with 9 years of flute-playing experience. Each was invited separately to perform the experiments.

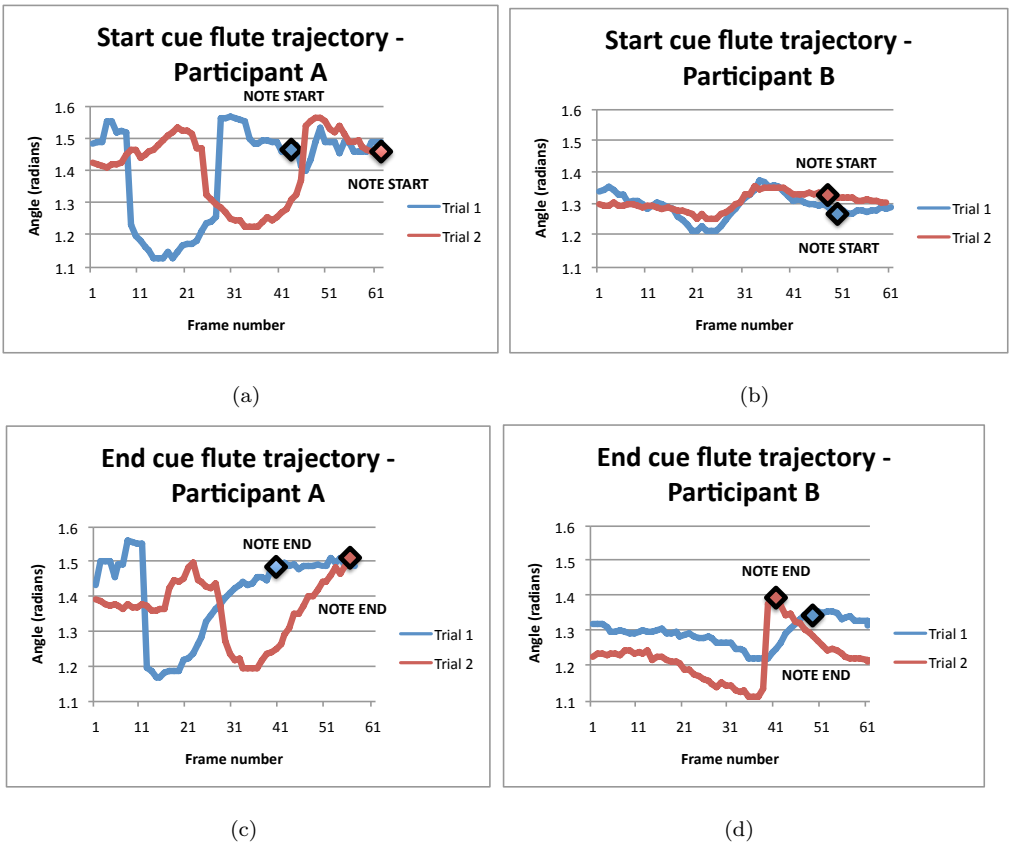
3.1.1 Gesture survey and analysis

In this experiment, we wanted to know whether flutists naturally used the gestures we hypothesized. That is, would they make start and end cues as we defined them, without prompting? The participant was given two sequences of duet music: one involving a simultaneous start (Fig. 9(a)), and one containing a fermata, requiring a simultaneous end of note (Fig. 9(b)). The participant was asked to play each musical sequence twice, assuming the role of leader. A secondary, advanced flute player familiar with the system assumed the role of follower, hereafter referred to the Follower. At no point did the participant receive any guidance as to how to lead. Their movements were filmed with a video camera at 25 fps for offline visual analysis and recognition by our system.

3.1.2 Gesture analysis

We plotted the angle of their flutes over time leading up to the start of a note (Figure 8(a) and 8(b)) and end of a note (Figure 8(c) and 8(d)); the lower the angle, the more the flute end is pointing downward, and so on. The audible beginnings and ends of notes have also been indicate with a diamond.

From these trajectories, we can validate our state machines for start and end gestures. Indeed, for



(a) (b)

(c) (d)

Figure 8: Resulting flute trajectories for a)-b) participants start cues and c)-d) participants end cues.

each start cue, we notice a DOWN-UP-DOWN trajectory before the note onset. In fact, Participant A's movements implies an additional state: UP-DOWN-UP-DOWN. However, Participant B's movement is not so consistently complex. The minimal sequence across our two players therefore appears to be DOWN-UP-DOWN. As for the end cue, we can also verify from the figures that there is a characteristic DOWN-UP motion before the end of the note, as hypothesized in Section 2.1.

A few other interesting points were noticed during this experiment. Firstly, Participant A's movements were much larger and pronounced that Participant B's. This implies that the method should be able to handle both large and small gestures. According to Wanderley et al.'s [27] study, this difference in magnitude may be expected: when they asked performers to perform with more exaggerated movements, they made the same movements, simply with a higher magnitude. Secondly, although not marked, a sharp breath intake sound could be heard before each note start. This breath sound is another indicator for start of play, as discussed in [44]. This may be a physiological correlation with the gesture, as the flute is raised when the player's lungs fill quickly with air. It is possible that the start gesture may not be purely iconic, but in fact be linked with the physical phenomenon of playing.

Through this experiment, we can check that the system is natural to use, without resorting to subjective surveys. Our system indeed was able to detect these four cues with a minimum state machine speed threshold $\Delta\theta = 0.003$ radians/frame (equivalent to 0.07 radians/sec given our frame rate of 25 fps). We used the speed threshold derived from this survey to evaluate our system in the next section.

3.1.3 Recognition rates

In this experiment, we set the $\Delta\theta = 0.003$ and asked each flutist to play the role of the leader for the music in Fig.7(c). We asked them to perform this excerpt five times with the Follower, and five times alone. Across our two participants, this resulted in ten start gestures and ten end gestures, for a total of 20 samples for each gesture type. Our system was able to detect all 20 of the gestures, with 3 false detections of start gestures. Indeed, a false start can be disastrous for a live performance, but this is also a challenge for musicians. As stated in a conducting technique guide [45]: "nothing before the start must look like a start. There must be no mystic symbols, twitches, or other confusing motions." Avoiding accidentally cue-ing a start is indeed a tough problem for humans as well as computational systems.

3.2 Experiment 2: Performance of audio-visual beat fusion module

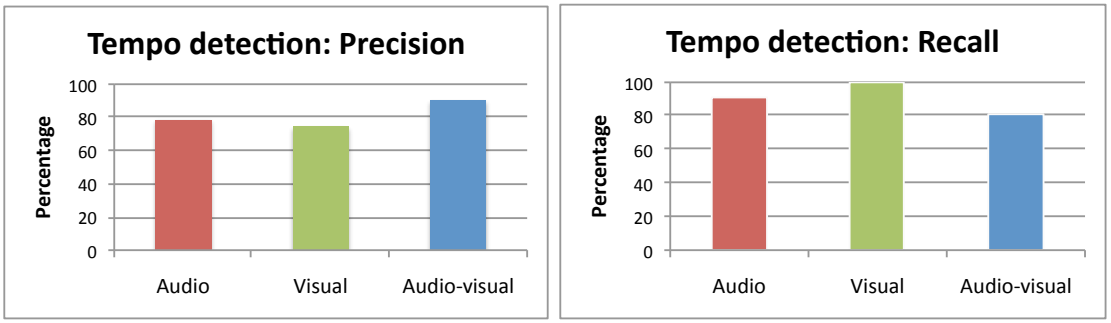
In this experiment, we evaluate our system's note onset and tempo detection accuracy.

A. Visual and Audio Beat Detections

An advanced flute player with 18 years of experience, equipped with a lapel microphone, played two legato notes in alternation, with no tonguing: A2 and B♭2 at approximately 66 bpm. With each change in note, the flutist performed a visual beat cue. A secondary observer, a classically trained intermediate-level clarinet player, tapped a computer key along with the changes in notes to provide a human-detected tempo (measured in IOI) for comparison.

The average absolute IOI error between our audio-visual tempo detection and the human detected tempo was 46 ms with a standard deviation of 32ms. On the other contrary, the relative IOI error (i.e. taking into account whether the error was negative and positive) was -1ms over 72 matched beats. This means that despite going too slow or fast during the piece, the robot would still end at virtually the same time as the human.

As for beat onset error, we found a mean of 180ms and a standard deviation of 47ms. The onset error was high, but not indicative of the system's performance; the groundtruth onsets were consistently tapped 100-200ms later than the system, possibly due to the human's motor delay compared to the audio. In Rasch's [46] synchronization experiments of wind and string instrument ensembles, asynchronization was defined by the standard deviation of onset time differences, and ranged from 20 ms for fast tempos to 50 ms for slower tempos. As our experimental tempo was relatively slow, the asynchronization of 47ms falls into a range comparable to human performers. Furthermore, as noted by Rasch, the smooth, relatively long rise time of wind and string instruments (20-100ms) allow for imperfectly aligned onsets to still be perceptually synchronized. Since the theremin also has a relatively indistinct, long rise time, we believe that this is an acceptable result.



(a) (b)

Figure 9: Beat tracking experiment results

B. Utility of audio-visual beat fusion module

In the final experiment, we verified the tempo estimation given by our audio-visual beat tracking system. We asked the two participants from Experiment 1 to play 8 notes with a ritardando, as shown in Figure 7(d). They were asked to perform this a total of ten times, using a gesture to keep in synchrony with their co-player. The first five times, this co-player was a real person, the Follower. This condition was used to give the flutist the context of a natural situation; it was essentially a "warm-up". The latter five times, we asked the flutist to imagine a co-player, but in reality play their notes alone. We used the latter condition to give us a total of 10 instances, 80 gestured notes, or 70 inter-onset-interval tempo indications. The rationale here was to ensure the algorithm was not be affected by the Follower's notes.

The result of the audio-visual tempo detection is shown in Figure 9. We can notice that beat fusion was better for precision, and worse for recall. In other words, our use of two sources of data prevented unwanted changes in tempo which could be disturbing for a musical performance. In summary, the system misses more tempo change signals than our vision-only approach, for example, but is robust to extraneous movements. This is likely preferable and somewhat similar to a human co-players' true behavior.

An unexpected outcome of this experiment came from the with-partner and alone conditions. Although the system could detect the cues in both conditions, a few times Participant A asked to stop and retry during the alone condition, citing she had made a mistake. We suggest that this was because of the visual feedback given by the Follower. Indeed, the physical synchronization of both players seemed to add to the ease of performing the movement. This is consistent with the "muscular bonding" phenomenon cited in the Introduction; synchronization of not only sound, but movement could be key. This implies that the robot should give some synchronized visual feedback, perhaps by head nodding.

4 DISCUSSION AND FUTURE WORK

Here, we summarize and discuss the results of this study.

4.1 Effectiveness of visual start and end cues

Visual cues are especially effective when no audio information is available. In this work, we made use of the movement that conductors make when showing musical beats; an up-and-down motion. This movement is also a natural way to express beats for most flutists, and we have shown that they are used for start and end cues too. If musical performers use similar gestures when they play other instruments, our method can be applied to a wide range of instruments.

4.2 Effectiveness of visual beat cues

Beat cues are provided in the middle of the ensemble performance. While these cues are considered to be good information when starting a new passage with a different tempo, it is yet to be confirmed whether these cues are appropriate when the human provides subtle tempo changes during his/her performance. Our future work includes the verification how the behavior of the robot is improved with employing an improved beat tracking method or a score following module.

4.3 Adaptation to tempo fluctuation

Because adaptation to tempo fluctuation is an inevitable issue to realize a human-robot ensemble, a robust beat tracking or score following method is necessary. We are currently seeking a score-following method based on a particle filter [47] [48]. The score following method produces better results compared to beat tracking methods. We are currently working to apply our visual detection modules to the score following method for a more robust co-player robot system.

Nevertheless, the score following method still suffers from the cumulative error problem; the error in the audio-to-score alignment accumulates due to tempo fluctuation. To cope with these cumulative errors, we need an error recovery mechanism at a higher level; e.g., the robot would jump to a certain passage when the robot detects a salient melody before the passage.

One of the drawbacks in our method is that the tempo detection accuracy depends on how skilled the flute player is. By using a larger matching threshold, we can suppress the false detection of beat cues. We need further experiments to determine the proper threshold.

4.4 Perspective on information fusion

Aggregation of visual and audio information is categorized into three methods:

1. **Visual information is used to filter audio information.** We have presented this type of information aggregation: among the detected audio onsets, some audio onsets irrelevant to the visual cue are filtered out so as to stabilize the tempo estimation accuracy.

2. **Audio is used to filter visual information.** Shiratori et al. uses the audio to segment dancing motions in a video [49]: Among the detected pose candidates for the dancing segmentation, audio beats are used to filter out false-detected poses.

3. **Both audio and vision are used equivalently.** Itohara et al.'s beat tracking method [48] uses the trajectory of guitar-playing arm motion and the audio beats in the guitar performance. These two information sources are integrated by a particle filter to obtain the improved tempo estimation and beat detection.

In addition, our current framework is only useful for slow and sparsely notated musical pieces when setting a large window to filter the audio beats. This may be useful for these cases, since it has been shown that synchronization is most difficult for slow pieces [46]. On the other hand, skilled musicians tend to play fast passages without any unnecessary motions [27]. We need a mechanism to robustly estimate the tempo when a fast and densely notated phrase is given as an input with little visual information like gestures.

Other remaining issues include the fact that the robot is only following the human leader. For true interaction, the human should also react to the robot's actions. Additionally, the robot should have its own their internal timing controller; for instance, Mizumoto et al. [50] employs an oscillator model to synchronize not only the tempo, but the phase of beat onsets. Other future directions include experiments with an augmented number of subjects, the use of robot-embedded microphones, and extension of the system to other instruments.

5 CONCLUSION

Our ultimate goal is to create a robot that can play music with human-like expressiveness and synchronicity, for better human-robot symbiosis. In this paper, we have developed a singing, theremin-playing robot that can synchronize in timing and speed with a co-player. Our novel contribution is the addition of visual cues for beat-tracking; we show that the system can estimate a flutist's tempo quickly, and with better robustness than with audio alone. We have also validated our hypothesized flute gesture trajectories with a small-scale experiment, suggesting that the robot can detect naturally-occurring cues.

6 ACKNOWLEDGMENTS

This paper is an extended version of the paper [24]. This work was supported by a Grant-in-Aid for Scientific Research (S) (No. 19100003), a Grant-in-Aid for Scientific Research in Innovative Areas (No. 22118502) and the Global COE program.

REFERENCES

[1] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. Koay, and I. Werry, "What is a robot companion-Friend, assistant or butler?," in *IROS*, Edmonton, pp. 1192–1197, 2005.

[2] H. Mizoguchi, T. Sato, K. Takagi, M. Nakao, and Y. Hatamura, "Realization of expressive mobile robot," in *ICRA*, Albuquerque, pp. 581–586, 1997.

[3] S. Brown and U. Volgsten, *Music and manipulation: On the social uses and social control of music*. Berghahn Books, 2006.

[4] S. Wiltermuth and C. Heath, "Synchrony and cooperation," *Psychological Science*, vol. 20, no. 1, pp. 1–5, 2009.

[5] W. H. McNeill, *Keeping together in time: dance and drill in human history*. Harvard University Press, 1995.

[6] E. Cohen, R. Ejsmond-Frey, N. Knight, and R. Dunbar, "Rowers ' high : behavioural synchrony is correlated with elevated pain thresholds," *Biology Letters*, vol. 6, no. 1, pp. 106–108, 2010.

[7] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of nonverbal behavior*, vol. 27, no. 3, pp. 145–162, 2003.

[8] P. M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-gruber, and F. Ric, "Embodiment in Attitudes, Social Perception, and Emotion," *Personality and Social Psychology Review*, vol. 9, no. 3, pp. 184–211, 2005.

[9] T. Mizumoto, A. Lim, T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H. Okuno, "Integration of flutist gesture recognition and beat tracking for human-robot ensemble," in *IROS Workshop on Robots and Musical Expressions*, Taipei, 2010.

[10] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?.," *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.

[11] C. Raphael, "A Bayesian network for real-time musical accompaniment," in *NIPS*, Vancouver, pp. 1433–1439, 2001.

[12] R. Dannenberg, "An on-line algorithm for real-time accompaniment," in *ICMC*, Paris, pp. 193–198, 1984.

[13] T. Otsuka, T. Takahashi, H. Okuno, K. Komatani, T. Ogata, K. Murata, and K. Nakadai, "Incremental polyphonic audio to score alignment using beat tracking for singer robots," in *IROS*, St. Louis, pp. 2289–2296, 2009.

[14] B. Vercoe and M. Puckette, "Synthetic rehearsal: Training the synthetic performer," in *ICMC*, Vancouver, pp. 275–278, 1985.

[15] G. Weinberg and S. Driscoll, "Robot-human interaction with an anthropomorphic percussionist," in *SIGCHI*, Montreal, pp. 1229–1232, 2006.

[16] K. Murata, K. Nakadai, R. Takeda, H. Okuno, T. Torii, Y. Hasegawa, and H. Tsujino, "A beat-tracking robot for human-robot interaction and its evaluation," in *Humanoids*, Daejeon, pp. 79–84, 2008.

[17] J. W. Davidson and A. Williamon, "Exploring co-performer communication," *Musicae Scientiae*, vol. 1, no. 1, pp. 53–72, 2002.

[18] L. De Bruyn, M. Leman, D. Moelants, and M. Demey. "Does social interaction activate music listeners?," *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music*, Copenhagen, pp. 93–106, 2009.

[19] K. Katahira, T. Nakamura, S. Kawase, S. Yasuda, H. Shoda, and M. Draguna, "The Role of Body Movement in Co-Performers' Temporal Coordination," in *ICoMCS*, Sydney, pp. 72–75, 2007.

[20] Werner Goebl and Caroline Palmer, "Synchronization of timing and motion among performing musicians," *Music Perception*, vol. 26, no. 5, pp. 427–438, 2009.

[21] W. E. Fredrickson, "Band musicians' performance and eye contact as influenced by loss of a visual and/or aural stimulus," *Journal of Research in Music Education*, vol. 42, no. 4, pp. 306–317, 1994.

[22] T. Mizumoto, H. Tsujino, T. Takahashi, T. Ogata, and H. G. Okuno, "Thereminist robot : development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model," in *IROS*, St. Louis, pp. 2297–2302, 2009.

[23] A. Lim, T. Mizumoto, L.-K. Cahier, T. Otsuka, T. Takahashi, K. Komatani, T. Ogata, and H.G. Okuno. "Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist," in *IROS*, Taipei, pp. 1964–1969, 2010.

[24] A. Lim, T. Mizumoto, L.-k. Cahier, T. Otsuka, T. Ogata, and H. G. Okuno, "Multimodal gesture recognition for robot musical accompaniment," in *RSJ*, Nagoya, 2010.

[25] G. Luck and J. A. Sloboda. "Spatio-temporal cues for visually mediated synchronization." *Music Perception* vol. 26, no. 5, pp. 465–473, 2009.

[26] T. M. Nakra, "Synthesizing Expressive Music Through the Language of Conducting," *Journal of New Music Research*, vol. 31, no. 1, pp. 11–26, 2002.

[27] M. Wanderley, B. Vines, N. Middleton, C. McKay, and W. Hatch, "The musical significance of clarinetists' ancillary gestures: an exploration of the field," *Journal of New Music Research*, vol. 34, no. 1, pp. 97–113, 2005.

[28] D. Overholt et al., "A multimodal system for gesture recognition in interactive music performance," *Computer Music Journal*, vol. 33, 2009, pp. 69-82.

[29] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[30] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[31] R. C. Bolles and M. A. Fischler, "A RANSAC-based approach to model fitting and its application to finding cylinders in range data," in *IJCAI*, Vancouver, pp. 637–643, 1981.

[32] Lim et al., "Robot Musical Accompaniment: Integrating Audio and Visual Cues for Real-time Synchronization with a Human Flutist", in *IPSJ*, Tokyo, 2010

[33] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[34] Schloss, *On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis*. PhD thesis, Stanford, CA, 1985.

[35] J. Bello, "Phase-based note onset detection for music signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, p. 49, 2003.

[36] C. Duxbury, J. Bello, and M. Davies, "Complex domain onset detection for musical signals," in *DAFx*, London, pp. 1–4, 2003.

[37] P. M. Brossier, *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Queen Mary University of London, 2006.

[38] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *ICMC*, Singapore, pp. 163–166, 2003.

[39] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System 'HARK': Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol. 24, no. 23, pp. 739–761, 2010.

[40] D. Mills, "Network Time Protocol (Version 3) specification, implementation and analysis," 1992.

[41] R. B. Dannenberg, S. Cavaco, E. Ang, I. Avramovic, B. Aygun, J. Back, E. Barndollar, D. Duterte, J. Grafton, R. Hunter, C. Jackson, U. Kurokawa, D. Makuck, T. Mierzejewski, M. Rivera, D. Torres, and A. Yu, "The Carnegie Mellon Laptop Orchestra," in *ICMC*, Copenhagen, pp. 340–343, 2007.

[42] H. Kenmochi and H. Ohshita, "VOCALOID – Commercial singing synthesizer based on sample concatenation," in *Interspeech*, Antwerp, pp. 4011–4010, 2007.

[43] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[44] H. Yasuo, I. Ryoko, N. Masafumi, and I. Akira, "Investigation of Breath as Musical Cue for Accompaniment System," *IPSJ SIG Technical Reports*, vol. 2005, no. 45(MUS-60), pp. 13–18, 2005.

[45] B. McElheran, *Conducting technique: for beginners and professionals. Oxford University Press, USA, 2004.*

[46] R. A. Rasch, Timing and synchronization in ensemble performance in: J. Sloboda (Ed.) Generative Processes in Music. Oxford University Press, 1988.

[47] T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H.G. Okuno, "Real-Time Audio-to-Score Alignment using Particle Filter for Co-player Music Robots", *EURASIP Journal on Advances in Signal Processing*, vol. 2011.

[48] T. Itohara, T. Mizumoto, T. Otsuka, T. Ogata, H. G. Okuno, "Particle-filter Based Audio-visual Beat-tracking for Music Robot Ensemble with Human Guitarist", in IROS, San Francisco, 2011, accepted.

[49] T. Shiratori, *Synthesis of dance performance based on analyses of human motion and music*, Ph.D. Thesis, University of Tokyo, 2006.

[50] T. Mizumoto, T. Otsuka, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, H. G. Okuno, "Human-Robot Ensemble between Robot Thereminist and Human Percussionist using Coupled Oscillator Model", in *IROS*, Taipei, pp. 1957–1963, 2010.