2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Who is the leader in a multiperson ensemble? — Multiperson human-robot ensemble model with leaderness —

Takeshi Mizumoto, Tetsuya Ogata, and Hiroshi G. Okuno

Abstract— **This paper presents a state space model for a multiperson ensemble and an estimation method of the onset timings, tempos, and leaders. In a multiperson ensemble, determining one explicit leader is difficult because (1) participants' rhythms are mutually influenced and (2) they compete with each other. Most ensemble studies however assumed that one leader exists at a time and the others just follow the leader. To deal with the multiple and time-varying leaders, we define *leaderness* indicating the power to influence the others as the product of the tempo stability and the distance from the ensemble tempo. This definition means that a leader should have a strong desire to change the current tempo. Using the leaderness, we present a state space model of a multiperson ensemble and an unscented Kalman filter based estimation method. The model consists of the leaderness update, the ensemble tempo update, the individual tempo update, and the onset timing adaptation, each of which has a relationship to psychological results of an ensemble. We evaluate our method using simulation and human behavior. The simulation results show that our model is stable for various initial tempos and the number of participants. For the human behavior, pairs and triads of participants are asked to tap keys in synchronization with the others. The results show that the leaderness successfully indicate the dynamics of the leaders, and the onset errors are 181msec and 241msec for pairs and triads on average, respectively, which are comparable to those of humans (153msec and 227msec for pairs and triads, respectively.)**

I. INTRODUCTION

Ensembles of people are common to most societies. The number of ensemble co-players can range from two to dozens, e.g., from a duo ensemble to an orchestra. In each ensemble, people can synchronize their playing timings each other. The purpose of our study is to enable the robots to play their instruments in synchronization with co-players in such multiperson ensembles.

In a multiperson ensemble, the leaders have two features:

1) **Multiple leaders exist**: For example, if the ensemble score has two main melodies, they will be the leaders, and the bass part and the drummer will cooperate to lead the ensemble, i.e., the bass part and the drummer are the leaders.
2) **Leader transition occurs**: For example, if the current leader makes a mistake, another co-player will be a new leader to keep the rhythm, and when a main melody part in a score changes, the leader will change.

Because of these features, the ensemble co-players have to decide who to follow by playing their scores. Conventional human-robot ensemble studies avoided the problem by assuming that only one leader exists at the same time [1]–[3]. This assumption is also used in a tapping task, which is used to investigate human's response to rhythms in psychology; the leader is the stimulus and the follower is a participant.

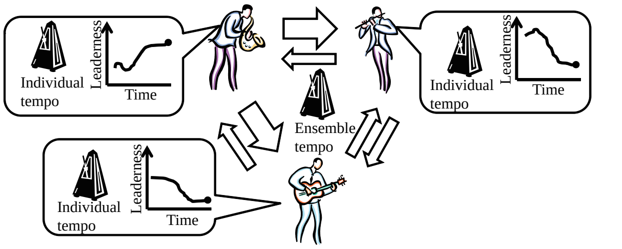


Fig. 1. Time-varying and continuous leaderness: Each participant has an individual tempo and leaderness. A high leaderness means a high power to influence the others. These tempos are aggregated into the ensemble tempo.

We present a method for estimating the multiple and time-varying leaders and rhythm in a multiperson ensemble that is based on a state space model. We define the ensemble tempo, i.e., the consensus of the tempo shared among all participants, as the aggregation of their individual rhythms. This is analogous to the multiperson decision making (MPDM) problem [4], which includes three steps: opinions with preference are given by experts, these opinions are aggregated, and they are modified in accordance with the aggregation. When the opinions converge, the experts have reached a consensus. MPDM problem has been widely studied, e.g., aggregation of different representations of preference [5] and an analysis of opinion dynamics [6]. We apply MPDM to a multiperson ensemble: an opinion is an individual tempo, they are aggregated into an ensemble tempo, and they are modified in accordance with the ensemble tempo.

The key idea is to define a *leaderness* value for each participant that quantifies his or her power to influence the others. Fig. 1 illustrates the ensemble with leaderness. We define it as the product of stability and distance from the ensemble tempo. Stability reflects the reliability of the participant, which is important because the tempo in an ensemble normally does not change frequently. Distance from the ensemble tempo reflects the participant's desire to change the tempo. Intuitively, a participant with a strong desire to change the ensemble tempo will have a higher leaderness. The leaderness distribution depends on a situation; if none of the participants has a desire to change and they simply try to match each other, i.e., no leaders exist, the leaderness is uniformly distributed and the ensemble tempo becomes the mean tempo. In contrast, if a participant tries to change the tempo, the leaderness is sparsely distributed and the

T. Mizumoto, T. Ogata, and H. G. Okuno are with the Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan. {mizumoto, ogata, okuno}@kuis.kyoto-u.ac.jp

978-1-4673-1735-1/12/$31.00 ©2012 IEEE

ensemble tempo will approach his or her tempo. From the viewpoint of a MPDM problem, leaderness corresponds to the expert's preference because, if he or she has a high leaderness, the ensemble tempo approaches his or her tempo.

On the basis of this idea, we build a nonlinear state space model consisting of three parts: a consensus model for updating the ensemble and individual tempos using the leaderness, a coupled oscillator model for adjusting the onset timing, and a procedure for updating the leaderness. This model has a relationship to psychological models, as discussed in section III. We estimate a hidden state using observed onset timings and inter-onset intervals (IOIs) based on an unscented Kalman filter (UKF) [7].

We review related works in Section II, present a multiperson ensemble model in Section IV, and discuss its relationship to psychological studies in Section III. We describe our evaluation of our method using simulation and a human's behavior in Section V. We conclude in Section VI with a brief summary and a mention of future work.

II. STATE-OF-THE-ART ENSEMBLE STUDIES

Musical ensembles have been actively studied from various points of view, such as human-robot interaction, psychology, and neuroscience. We review these studies here as we build a human ensemble model for human-robot interaction.

Human-robot ensembles using natural modalities such as visual and audio cues have been studied by assuming that an ensemble has one leader at a time. Many duo ensembles have been developed by assuming a human leader and a robot follower, e.g., a thereminist robot follows a drummer by predicting the onset [1], a flutist by recognizing his or her gesture [8], and a guitarist by audio visual integration [9]. A flutist robot follows a saxophonist by recognizing the gesture and phrase [10]. An ensemble with more than two participants is also developed by Weinberg *et al.* [3]. This is a quartet ensemble among two humans and two robots. They treated the leader transition by using a turn-taking model, i.e., only one leader exists at a time.

A person's rhythm recognition has been studied from the psychological point of view. A tapping task, which is used in many studies, is based on the assumption that the participant is the follower and the stimulus, e.g., a metronome sound or a sequence of tones, are the leader. This is because a participant is instructed to tap a key by following the stimulus. The task is often used to build psychological models (e.g., Haken *et al.* [11] and Large *et al.* [12]). Some studies have investigated behavior using human interactions, however, these studies focused on the mental processes of individuals, not their interactions. For example, a free jazz improvisation was studied using a cybernetic model [13], and a duo ensemble of pianists was studied and found a correlation between a pianist's ability to image music from a score and a sensorimotor synchronization [14].

Our model uses an oscillator, i.e., a timekeeper, as a model of a participant's onset generation. The existence of a timekeeper for rhythm recognition and generation is a commonly agreed upon hypothesis; neurophysiological studies suggests the existence of a neural clock, i.e., timing representation using pulses or oscillators, in our brain (see [15], [16] for detailed reviews), and psychological studies explain human time keeping using oscillators [11], [12]. An oscillator is also used for music processing, such as onset prediction [1], robot drumming [17], and beat tracking [18].

TABLE I

NOTATIONS

t	Time
Δt	Time interval
N	Number of ensemble participants
M	Number of past onsets used to calculate stability
i	Index of an ensemble participant ($i \in \{1, ..., N\}$)
$\omega_s(t)$	Aggregated ensemble rhythm
$\omega_i(t)$	ith tempo
$\theta_i(t)$	ith phase (Onset is produced when $\theta_i(t) = 2n\pi$.)
$l_i(t)$	Leaderness of ith participant ($\sum_{i=1}^{N} l_i(t) = 1$)
$\mathbf{x}(t)$	State vector, $\mathbf{x}(t) \in R^{1+N(M+2)\times 1}$
$\mathbf{z}(t)$	Observation vector, $\mathbf{z}(t) \in \int R^{2N\times 1})$

III. PSYCHOLOGICAL MODELS OF ENSEMBLES

We introduce an ensemble psychological model proposed by Keller [19] because our model is related to the model. Relationship to our model is discussed after definition of our model in section IV-E. Briefly, his model of the cognitive process in an ensemble consists of three parts: anticipatory auditory imagery, prioritized integrative attention, and timing adaptation.

The anticipatory auditory imagery is a mental imagery of a participant, i.e., imaging the ideal music of an ensemble, which is a goal shared among all the participants. The imagery makes the player's effectors move in synchronization with the sound by anticipating the performances of the others. While playing, each participant modifies his or her image on the basis of the others' performances. Sensorimotor synchronization, i.e., an ability to generate a motion in synchronization with the heard sound, and the ability to making imagery have a positive correlation [14], [20].

The prioritized integrative attention is a process to pay attention to each participant with priority and integrate them with the priority. Because each participant has to maintain both the performance correctness and synchronization with the others, the attention needs to be paid to each of the others. The amount of the attention is not equal, more for the leaders and less for the followers. The prioritized integrative attention is a hybrid of selective and non-prioritized attentions: The former is a model to select only one leader, and the latter is a model to pay attention to all the others equally.

Timing adaptation is a process for synchronizing the beat. This process corresponds to the mental timekeeper discussed in Section II. To compensate for a mismatch in onset timings during an ensemble, the participants adjust their mental timekeepers. The correction is twofold: phase and period correction [21]. Phase correction is a timing update that occurs regularly and automatically whereas period correction is a tempo update that occurs only if the participant notices an obvious tempo change. This correction occurs not only for the beat timings but also for subdivisions of beats [22].

IV. MULTIPERSON ENSEMBLE MODEL

We define a multiperson ensemble problem, build a state space model including a state update model and an observation model, then, we describe UKF for state estimation.

A. *Problem Statement and Model Overview*

Problem statement

Input: Last onset timings and IOIs of all participants
Output:
Leaderness of all participants
Next onset timings and tempos of all participants
Assumptions:
(1) All participants generate onsets for each quarter note
(2) All participants try to synchronize with each other

The inputs can be estimated from beat tracking, e.g., [23] which is designed for music robots. The outputs are the next onset timings and tempos since the robot needs to predict the human's playing. Using the first assumption, we avoid treating a metrical structure. This assumption can be relaxed to accept more complex structures by representing the individual rhythm as coupled oscillators [24]. The second assumption is obvious given that ensemble participants generally try to optimize their performances which require synchronization.

We solve this problem by building the state space model of participants' rhythms shown in Fig. 2, and by estimating the state from the observations using an UKF. The notations are summarized in Table I. In our model, the ith participant has an individual tempo $\omega_i(t) > 0$ and a phase $\theta_i(t) > 0$. The ensemble tempo $\omega_s(t)$ is updated by aggregating $\omega_1(t), ..., \omega_N(t)$ using $l_i(t), ..., l_N(t)$. The individual tempos are updated so that they converge to $\omega_s(t)$ with a step size of $l_i(t)$. $\theta_i(s)$ is updated using both $\omega_i(t)$ and the phase differences for each other. $l_i(t)$ is updated using $\omega_s(t)$ and the past M tempos, $\omega_i(t), ..., \omega_i(t-M-1)$. When $\theta_i(\tau) = 2n\pi (n \in \mathbf{N})$, the system observes the ith onset time τ and tempo $\omega_i(\tau)$.

B. *State Update Model*

The state update model consists of three parts; tempo, phase, and leaderness.

1) Tempo: If a participant has a higher leaderness, his or her tempo should have a greater effect on the ensemble tempo. The ensemble tempo is then aggregated using the leaderness-weighted average of the individual tempos:

$$\omega_s(t+1) = \sum_{i=1}^{M} l_i(t)\omega_i(t). \tag{1}$$

The participants adjust their tempos so that they converge to the ensemble tempo, and the amount of adjustment should less for a participant with a higher leaderness. This is because the leader plays the music as it is, and the followers follow the leader. Therefore, the update equation of $\omega_i(t)$ is:

$$\omega_i(t+1) = \omega_i(t) + (1 - l_i(t))(\omega_s(t) - \omega_i(t)). \tag{2}$$

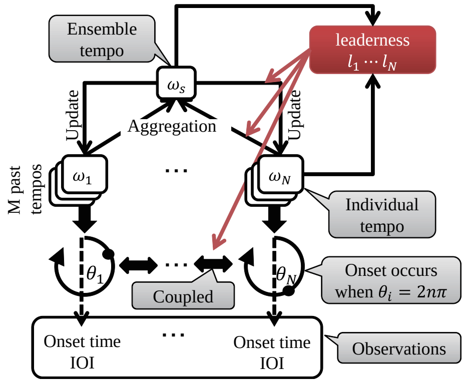


Fig. 2. Overview of multiperson ensemble model

Note that $1 - l_i(t)$ is the total of the others' leaderness because $\sum_{i=1}^{N} l_i = 1$ according to Table I.

This calculation corresponds to the ordered weighted average in a consensus model [4], which is a model of MPDM, e.g., policy making in a political committee. The opinions are aggregated into one opinion, and the experts modify their opinions slightly. These aggregation and modification processes are iterated until the opinions converge. In a multiperson ensemble, the participants are experts having opinion $\omega_i(t)$ and preference $l_i(t)$. Individual tempos are aggregated into a tentative consensus, $\omega_s(t)$, and modified in accordance with the consensus.

2) Phase: The beat generation of a participant is represented using a coupled oscillator model [1] because onsets are generated periodically and their timings affect each other. The coupled oscillator model consists of multiple oscillators having a velocity and phase. Each oscillator generates an onset when its phase becomes a multiplier of 2π. The phase increases in accordance with the tempo and is modified to minimize the phase difference. This difference is measured as a 2π periodic function called the coupling function, and the amount of adjustment is called the coupling strength.

We use a sinusoidal function as the coupling function, which is known as Kuramoto model [25], and the leaderness as the coupling strength. This is because the phase difference of a participant having higher leaderness should influence the others more. The phase update is formalized as

$$\begin{aligned} \theta_i(t+1) &= \theta_i(t) + \omega_i(t)\Delta t \\ &\quad + \sum_{j=1}^{N} l_j(t) \sin(\theta_j(t) - \theta_i(t)) \end{aligned} \tag{3}$$

3) Leaderness: We define the leaderness as the product of the stability and the distance from the ensemble tempo. This is because the leader should both have a desire to change the ensemble tempo and be stable enough to prevent the ensemble tempo from being perturbed by mistakes.

The stability, $s_i(t)$, of $\omega_i(t)$ is defined as the exponential of the standard deviation (std) of the past M tempos including the current tempo $\omega_i(t)$:

$$s_i(t) = \exp\left(-Std[\omega_i(t), \cdots \omega_i(t-M-1)]\right) \tag{4}$$

where $Std[x_1, ..., x_N]$ denotes an operator used to calculate the std. We define $s_i(t)$ as an exponential of std to limit the range of $s_i(t)$ in $[0, 1]$. Thus, $s_i(t)$ is maximum if the past M tempos are exactly the same and is less than the maximum if the tempo fluctuates.

The distance from the ensemble tempo, $p_i(t)$, is defined as the sigmoid-like function of the absolute difference:

$$p_i(t) = \frac{2}{3}\frac{1}{1+\exp\left(-\left|\omega_i(t)-\omega_s(t)\right|\right)} - \frac{1}{2}. \tag{5}$$

This definition is designed so that $p_i(t) \in [0, 1]$.

Finally, the leaderness is updated as

$$l_i(t+1) = s_i(t)p_i(t). \tag{6}$$

$l_i(t+1)$ is normalized to satisfy $\sum_{i=1}^{N} l_i(t+1) = 1$.

The leaderness determines the attractor of the individual tempos because they converge to the ensemble tempo (Eq. (2)), which is calculated using the leaderness (Eq. (1)). If the leaderness remains stable, the individual rhythms converge to the ensemble rhythm according to Eq. (2). Therefore, our definition can be interpreted as the leader being the participant who changes the tempo attractor.

4) State Vector: In summary, the state vector $\mathbf{x}(t)$ is

$$\begin{aligned}\mathbf{x}(t) \;\; = \;\; & (\omega_s(t), \omega_1(t), ..., \omega_N(t), \\ & \omega_1(t-1), ..., \omega_N(t-M-1), \\ & l_1(t), ..., l_N(t), \theta_1(t), ..., \theta_N(t))^T . \end{aligned} \tag{7}$$

$\mathbf{x}(t)$ has $1 + N(M+2)$ dimensions: the ensemble tempo (1 dim., Eq. (1)), the individual tempo (N dim., Eq. (2)), the individual phase (N dim., Eq. (3)), the leaderness (N dim., Eq. (6)), and the past $M-1$ tempos ($N(M-1)$ dim.).

C. Observation Model

In consideration of the problem statement in Section IV-A, the system can observe the participants' tempo and onset timings not every time interval but only at their onset timings. Therefore, the observation occurs partially since the onsets are not perfectly synchronized, and has an interval since an onset is generated only at the beat time.

We now design the observation model. Let the ith ($i = 1, .., N$) element of $\mathbf{z}(t)$ be $z_i(t)$. The first half elements of $\mathbf{z}$ are tempos and the rest of them are phases. When an onset from ith participant is not observed, we substitute $\emptyset$ to z_i and z_{i+N} to indicate no observation. When it is observed, we obtain two kinds of information; the participant's tempo (z_i) and the participant's phase being zero at the time (z_{i+N}).

Therefore, the observation model is:

$$z_i(t) \;\; = \;\; \begin{cases} \omega_i(t) & \text{if } \theta_i(t) = 2n\pi \\ \emptyset & otherwise \end{cases} \tag{8}$$

$$z_{i+N}(t) \;\; = \;\; \begin{cases} 0 & \text{if } \theta_i(t) = 2n\pi \\ \emptyset & otherwise \end{cases} \tag{9}$$

D. Unscented Kalman Filter for State Estimation

We estimate the state using an UKF [7] because both the state update and observation models are nonlinear. The UKF utilizes an unscented transform, which estimates the mean and covariance of a normal distribution after any nonlinear transformation from the given of mean and covariance before the transformation using deterministically selected samples. The UKF approximates the nonlinear state space model at least second order.

The observation vector can have $\emptyset$ because the observation is obtained partially at an interval. In that case, we skip the state estimation update with an observation, i.e., the innovation is assumed to be zero.

Let the robot be the jth participant. If the UKF successfully estimate the hidden state, the robot can predict when it should generate an onset from $\omega_j(t)$ and t. The robot can play an instrument based on this prediction.

Although no noise terms are included in IV-B and IV-C, we consider additive Gaussian noise which are the design parameters of the UKF.

E. Relationship to Psychological Model

Here we discuss the relationship between our model and the three parts of Keller's model.

The anticipatory auditory imagery, which is used for anticipation and is adjusted during the ensemble, corresponds to the tempo, $\omega_i(t)$ and $\omega_s(t)$ because of two reasons: (1) $\omega_i(t)$ is used for anticipating the next onset time because the system can predict ith participant's next onset time from the last one and the current tempo, $\omega_i(t)$, and (2) the participants adjust their tempos $\omega_i(t)$ so that they converges to $\omega_s(t)$ shown in Eq. (2). $\omega_s(t)$ is the shared goal of the ensemble because all $\omega_i(t)$s converge to $\omega_s(t)$.

The prioritized integrative attention corresponds to the leaderness $l_i(t)$, i.e., the priority of the attention. A participant with a higher leaderness receives more attention because $l_i(t)$ determines the influence of the ith participant, shown in Eqs. (1), (2), and (3). In addition, $l_i(t)$ is a hybrid of selective and non-prioritized attention; if the leaderness is sparsely distributed, i.e., only one participant has a value of one and those of the others are zero, the attention is selective because only one participant is recognized as the leader. If the value is equally distributed, i.e., $l_i(t) = 1/N$, the attention is non-prioritized because all participants equally influence each other.

The timing adaptation corresponds to the coupled oscillator model in Eqs. (2) and (3). The phase correction is modeled in Eq. (3) using the Kuramoto model. The period correction is modeled in Eq. (2).

V. EXPERIMENTS

We evaluate our model and estimation method using simulation and a multiperson tapping task.

A. Experiment 1: Simulation

In simulation experiment, we evaluate three aspects of our model and estimation method: (1) the convergent IOIs for different initial tempos, (2) the convergent IOIs for different number of participants, and (3) the onset estimation error for different number of participants. The first two experiments are designed for investigating the model, and the third one is designed for evaluating the estimation performance.

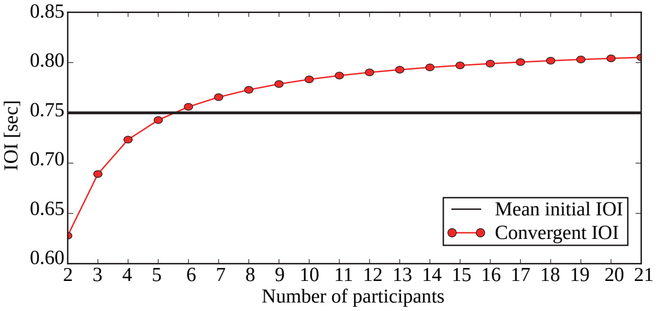


Fig. 3. IOIs for various numbers of participants: the black and red lines denote the mean of initial IOIs and the convergent IOIs.

1) *Settings:* These two parameters are commonly used; the time step Δt is 0.05sec, and the simulation is performed for 40sec. For the first experiment, we set $N = 3$, and the initial tempos $\omega_i(0)$ to be all the combinations of four different IOIs, 1.0, 0.75, 0.6, 0.5sec corresponding to 60, 80, 100, 120bpm, respectively. Therefore, the experiment is performed with $4^3 = 64$ conditions. For the second experiment, we use the number of participants N ranging from 2 to 21. The initial tempo of the ith participant is defined as $(60 + 60i/N)$bpm. In other words, the initial IOIs are equally spaced between 0.5 and 1.0sec. For the third experiment, we use the same conditions as the second experiment for model update. The number of past tempos M is set to 10, and the state update and observation noise matrices for UKF are the diagonal matrix whose diagonal component is 0.05.

The onset error is defined as the mean of the absolute difference between the nearest onset timings. Let $o_A^{(i)}$ be the ith onset time of a participant A, $o_B^{(j)}$ be the jth onset time of a participant B, and L be the number of A's onsets. The onset error e_{AB} is defined as

$$e_{AB} = \frac{1}{L}\sum_{i=1}^{L}\min_j \left| o_A^{(i)} - o_B^{(j)} \right| \qquad (10)$$

2) *Results and Discussion:* The first result is the convergent IOI of various initial tempos. Our state update model is stable because all three IOIs are converged to the same IOI after less than 500msec, for every combination. The difference between the mean of initial IOIs and the convergent IOI was 16msec on average with the std 25msec. This suggests that the mean of initial IOIs can be used for predicting the convergent IOI. This result is discussed in Section V-B.

The second result is the convergent IOIs for various numbers of participants shown in Fig. 3. When the number of participants is small (>5), the convergent IOI is less than the mean of initial IOI, and it increases as the number of participant increases. Intuitively, the convergent tempo becomes slower as the ensemble size becomes larger if no conductor exists.

The third result is obtained by the absolute onset estimation error. The mean of all number of participants was 120msec and the std was 36msec. The median of them was 80msec and the std was 40msec. The result that the mean is larger than the median suggests that the onset prediction succeeds mainly, but it fails occasionally.

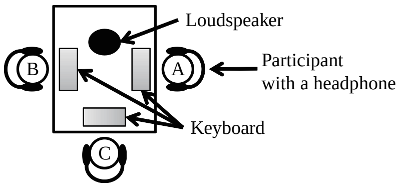


Fig. 4. Configuration: A, B, and C is the positions of the participants with headphones. Keyboards are placed in front of them. A beep sound is played through a loudspeaker, and a target tempo is played through a headphone. For the pair tapping task, C is not used.

TABLE II

STIMULUS FOR MULTIPERSON TAPPING EXPERIMENT

	Pre-stimulus			Main stimulus				
A	60	s_5	cue	s_{25}	s_{25}	80	s_{25}	cue
B	80	s_5	cue	s_{25}	50	s_{25}	s_{25}	cue

	Pre-stimulus			Main stimulus					
A	50	s_5	cue	s_{25}	80	s_{25}	s_{25}	s_{25}	cue
B	60	s_5	cue	s_{25}	s_{25}	50	s_{25}	s_{25}	cue
C	80	s_5	cue	s_{25}	s_{25}	s_{25}	80	s_{25}	cue

The top is for pairs and the below is for triads. s_5 and s_{25} denote the 5sec and 25sec of silence, respectively. The number denotes the target tempo in bpm given by a headphone. The cue denotes the 0.1sec, 880Hz pure tone.

B. *Experiment 2: Multiperson Tapping*

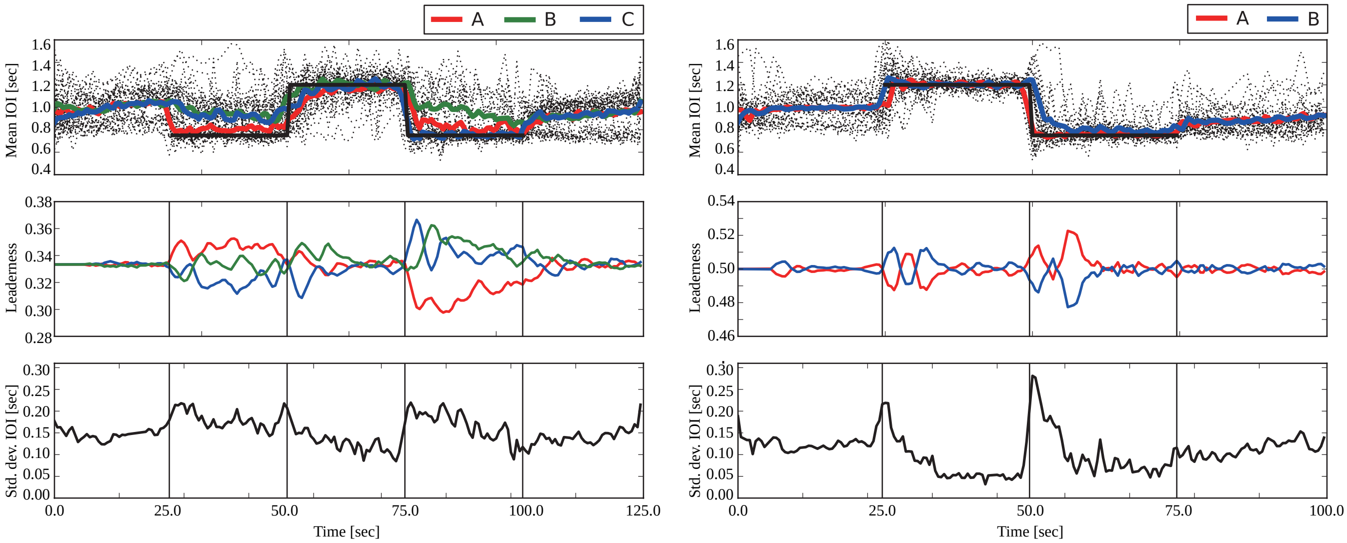
In this experiment, we analyze the human's behavior in a multiperson tapping task and estimation error.

1) *Settings:* We randomly formed four pairs (A and B) and three triads (A, B, and C) from nine participants without any motor or sensory impairment, ranging in age from 21 to 38 (8 men and 1 woman).

Each participant sat on a chair and was given a keyboard and a headphone (Fig. 4). When a participant tapped a key of a keyboard, a beep sound (880, 440, and 220Hz for participant A, B, and C, respectively) was played through a loudspeaker. Three instructions were given: (1) Initial tapping tempo is given at the beginning of each trial. Start tapping with the tempo when a start cue is heard from the headphone. (2) When no sound heard from the headphone, tap keys regularly and synchronously with others without looking at each other to avoid audio-visual integration. (3) When a metronome sound heard from the headphone, follow it and ignore the other tapping sounds.

The different stimuli are given to all participants as summarized in Fig. II. They were generated in advance and played using a multichannel audio player, Roland UA-101. The stimuli are twofold: pre-stimulus and main stimulus. In the pre-stimulus, different initial tempos were given for 10 seconds. After 5 seconds of silence, a start cue was given simultaneously. In the main stimulus, a metronome sound or a silence was given according to Fig. II. The duration of each column was 25sec. We expected that a participant given a metronome sound had a high leaderness. Finally, the trial was finished by the end cue.

For each pair or triad, we asked to do four trials, one for a practice and three for experiments. The durations of the trials were 115.2sec and 140.2sec for pairs and triads, respectively.



(a) Triad tapping

(b) Pair tapping

Fig. 5. Multiperson tapping result of triads and pairs: The top figure shows the IOI trajectories. The black solid line denotes the given tempo. The red, green, blue lines denote the mean trajectories of participants A, B, and C, respectively. The dotted black lines denote all trajectories. The middle figure shows the leaderness trajectory of three participants. The colors are the same as the top. The bottom figure shows the std of all trajectories. The vertical lines of the middle and the bottom figures denote the timing of the metronome tempo change.

2) Results and Discussion: Figs. 5(a) and 5(b) shows the result. Each panel shows the trajectories of mean IOIs, leaderness, and std of IOIs, respectively. Fig. 5(a) is divided into five parts, whose durations are 25sec, on the basis of the target IOI: no target, 0.75sec to A, 1.2sec to B, 0.75sec to C, and no target. Fig. 5(b) is also divided into four parts with the same duration: no target, 1.2sec to B, 0.75sec to A, and no target. The leaderness is smoothed by a moving average with window size 5 to discuss the trend. The leaderness is equally distributed and not updated until more than M samples are accumulated.

IOI trajectories

The top of Figs. 5(a) and 5(b) shows the tempo trajectories. Focusing on the mean trajectories of the participants and the target IOI (solid lines) of both figures, we can confirm that the participants did the task well. This is because when the target IOI is given to a participant, his or her IOI quickly approaches the target and those of the others approach later.

Focusing on the first and the last part of all trials (dotted lines) of both figures, we can confirm the dependency of convergent IOI on the initial tempo as we hypothesized in the simulation. This is because the dotted lines are widely spread in the first part whereas they spread narrower in the last one. This difference can be caused by the different initial tempos; the initial tempos are different for each trial because they need to maintain a given tempo for five seconds in their mind which is not precise. In contrast, the participants have similar IOIs through interaction before the last no-target part. Therefore, the trajectories are widely spread in the first part, and are narrow in the last part.

As shown in the last part in Fig. 5(a), one triad keeps the previous IOI whereas those of the others are increasing. This suggests that the members of the former triad have a high capability of time keeping. This dynamics difference depending on participants shows a limitation of our model because our model does not incorporate such an individual-dependent property.

Dynamics of leaderness

The middle of Figs. 5(a) and 5(b) show the leaderness dynamics. First, we discuss about Fig. 5(a). In the first and the last part, the leaderness is almost uniformly distributed. Since this means that the tempos among subjects are synchronized, the ensemble has no leaders in this time. In the second part, the leaderness of A is the highest and that of B and C are low, meaning that A is the leader, and B and C are the followers. In the third part, the B becomes the leader, however, the explicit leader disappears in the last of this part because the tempos have been converged as shown around 60sec in the top figure. In the fourth part, the C becomes the leader and A becomes the follower. The B still have a high leaderness because B keeps a tempo far from $\omega_s(t)$.

Next, we discuss about Fig. 5(b). The leaderness is uniformly distributed in the first and the last part, which are similar to those of Fig. 5(a). In the beginning of the second and third part, the leaderness is biased to the participant given a target tempo. In this case, the bias quickly disappears because of the quick IOI convergence. This indicates that the pair tapping task is easier than the triad one.

Standard deviation

The bottom of Figs. 5(a) and 5(b) shows the tempo convergence. The std of IOIs is high at the beginning of the all parts, then, it decreases over time. This indicates that the IOIs converge to the ensemble tempo, as in Eq. (2).

Onset and IOI errors

The onset errors among participants are summarized in Table III. The onset error definition is the same as Eq. (10) for the pair tapping, and is the mean of e_{AB}, e_{BC}, and e_{CA} for the triad tapping. Comparing the average errors of them, we find

TABLE III

Onset errors in multiperson tapping task

	Pair tapping		Triad tapping		
Trial	Mean	Std	Mean	Std	
1	123	53	247	60	
2	147	101	230	56	
3	188	50	204	66	
Average	153	68	227	61	[msec]

TABLE IV

Mean and median of IOI and onset estimation errors

	IOI error		Onset error		
Condition	Median	Mean	Median	Mean	
Pair tapping	33	63	181	204	
Triad taping	110	154	241	246	
Average	72	109	211	225	[msec]

that the triad task was more difficult because the error of the pair is smaller than the triad. Although the trajectories are not shown, the errors were especially large when the target IOI was slow. This is because predicting the onset is difficult in the slow tempo since the motion becomes less rhythmic. Although the error increases as do the trial more for the pair tapping whereas it decreases for the triad tapping, this is insufficient to discuss because the number of participants and trials are not enough.

The IOI and onset estimation errors of UKF are summarized in Table IV. The IOI error is the mean absolute error, and the onset error is the same as Eq. (10). The median IOI errors are 33msec for the pairs, and 110msec for the triads. The error in triad tapping is larger, which is the same as humans. The median onset errors are 181msec for the pairs, and 241msec for the triads, i.e., a robot controlled by our method will have these onset errors in ensemble. Recalling that the onset errors among humans were 153msec for the pairs and 227msec for the triads, the robot can play to a comparable accuracy with humans.

VI. Conclusion

We presented a state space model and its estimation method of multiperson ensemble for human robot ensemble. Our model has a relationship to psychological model in ensemble, and can be estimated from beats using the UKF. The main contribution is that our method realized the multiple and time-varying leaders using the leaderness quantifying the leadership in an ensemble. Experimental results showed that our model is stable for wide range of number of participants and initial tempos, the dynamics of the leaderness quantifies the changing leadership, and our method can predict the human's behavior in multiperson tapping.

We have three future plans; implementing our method on real human robot ensemble, integrating the visual information, i.e., modify the observation model, and introducing hierarchical oscillators to represent a complex rhythm.

Acknowledgments

We are grateful to Keisuke Nakamura of Honda Research Institute, Japan Co., Ltd., for his helpful advice on state space modeling. This study was partially supported by Honda Research Institute Japan Co., Ltd., a Grant-in-Aid for Scientific Research (S) (No. 19100003), for Scientific Research on Innovative Areas (No. 22118502), and for JSPS Fellows (No. 23·6572), and the Global COE Program.

References

[1] T. Mizumoto *et al.* "Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model", in *Proc. IROS*, 2010, pp. 1957–1963.

[2] A. Lim *et al.* "Robot musical accompaniment: Integrating audio and visual cues for real-time synchronization with a human flutist", in *Proc. IROS*, 2010, pp. 1964–1969.

[3] G. Weinberg *et al.* "The creation of a multi-human, multi-robot interactive jam session", in *Proc. NIME*, 2009, pp. 70–73.

[4] R.R. Yager, "On ordered wighted averaging aggregation operators in multicriteria decision making", *IEEE Trans. on SMC*, vol. 18, no. 1, pp. 183–190, 1988.

[5] E. H.-Viedma *et al.* "A consensus model for multiperson decision making with different preference structures", *IEEE Trans. on SMC Part A*, vol. 32, no. 3, pp. 394–402, 2002.

[6] V D. Blondel *et al.* "On krause's multi-agent consensus model with state-dependent connectivity", *IEEE Trans. on Automatic Control*, vol. 54, no. 11, pp. 2586–2597, 2009.

[7] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation", *Proc. the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[8] A. Lim *et al.* "A musical robot that synchronizes with a co-player using non-v erbal cues", *Advanced Robotics*, vol. 26, no. 3-4, pp. 363–381, 2012.

[9] T. Itohara *et al.* "A multi-modal tempo and beat tracking system based on audio-visual information from live guitar performances", *EURASIP J. on Audio, Speech, and Music Processing*, 2012, doi:10.1186/1687-4722-2012-6.

[10] K. Petersen *et al.* "Musical-based interaction system for the Waseda Flutist Robot: implementation of the visual tracking interaction module", *Autonomous Robots J.*, vol. 28, no. 4, pp. 439–455, 2010.

[11] H. Haken *et al.* "A theoretical model of phase transitions in human hand movements", *Biological Cybernetics*, vol. 51, pp. 347–356, 1985.

[12] E. W. Large and M. R. Jones, "The dynamics of attending: How people track time-varying events", *Psychol Rev*, vol. 106, no. 1, pp. 119–159, 1999.

[13] J. Braasch, "A cybernetic model approach for free jazz improvisations", *Kybernetes*, vol. 40, no. 7, pp. 984–994, 2011.

[14] N. Pecenka and P. E. Keller, "Auditory pitch imagery and its relationship to musical synchronization", in *The Neurosciences and Music III: Disorders and Plasticity*, pp. 282–286. New York Academy of Sciences, 2009.

[15] C V. Buhusi and W. H Meck, "What makes us tick? functional and neural mechanisms of interval timing", *Nature Reviews Neuroscience*, vol. 6, pp. 755–765, 2005.

[16] R. J. Zatorre *et al.* "When the brain plays music: auditory-motor interactions in music perception and production", *Nature Rev Neurosci*, vol. 8, pp. 547–558, 2007.

[17] S. Shaal *et al.* "Nonlinear dynamical systems as movement primitives", in *Proc. Humanoids*, 2001.

[18] E. W. Large, "Beat tracking with a nonlinear oscillator", in *Working Notes of Intl. Joint Conf. on Artificial Intelligence Workshop on Artificial Intelligence and Music*, 1995, pp. 24–31.

[19] P. E. Keller, "Joint action in music performance", in *Enacting Intersubjectivity: A Cognitive and Social Perspective on the Study of Interactions*, pp. 205–221. IOS Press, Amsterdam, 2008.

[20] P. E. Keller, "Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles", *Music Perception*, vol. 28, no. 1, pp. 27–46, 2010.

[21] M. H. Thaut *et al.* "Multiple synchronization strategies in rhythmic sensorimotor tasks: phase vs period correction", *Biological Cybernetcis*, vol. 79, pp. 241–250, 1998.

[22] B. H. Repp and H. Jendoubi, "Flexibility of temporal expectations for triple subdivision of a beat", *Advances in Cognitive Psychology*, vol. 5, pp. 27–41, 2009.

[23] K. Murata *et al.* "A robot uses its own microphone to synchronize its steps to musical beats while scatting and singing", in *Proc. IROS*, 2008, pp. 2459–2464.

[24] E. W. Large and C. Palmer, "Perceiving temporal regularity in music", *Cognitive Sceince*, vol. 26, pp. 1–37, 2002.

[25] Y. Kuramoto, *Chemical oscillations, waves, and turbulence*, Dover Publications, 2003.