(Accepted for publication in Neural Networks, 2015)

Development of Compositional and Contextual Communicable Congruence in Robots by Using Dynamic Neural Network Models

Gibeom Park and Jun Tani[a,*]

Dept. of Electrical Engineering, KAIST, Yuseong-gu, Daejeon, Republic of Korea

*Corresponding author

Correspondence should be sent to J. Tani

Email address: tani1216jp@gmail.com

[a]Tel: +82-42-350-7428

[a]Postal address: Room 516, N1 Building, 291 Daehak-ro(373-1 Guseong-dong), Yuseong-gu, Daejeon 305-701, Republic of Korea

Abstract

The current study presents neurorobotics experiments on acquisition of skills for “communicable congruence” with human via learning. A dynamic neural network model which is characterized by its multiple timescale dynamics property was utilized as a neuromorphic model for controlling a humanoid robot. In the experimental task, the humanoid robot was trained to generate specific sequential movement patterns as responding to various sequences of imperative gesture patterns demonstrated by the human subjects by following predefined compositional semantic rules. The experimental results showed that (1) the adopted MTRNN can achieve generalization by learning in the lower feature perception level by using a limited set of tutoring patterns, (2) the MTRNN can learn to extract compositional semantic rules with generalization in its higher level characterized by slow timescale dynamics, (3) the MTRNN can develop another type of cognitive capability for controlling the internal contextual processes as situated to on-going task sequences without being provided with cues for explicitly indicating task segmentation points. The analysis on the dynamic property developed in the MTRNN via learning indicated that the aforementioned cognitive mechanisms were achieved by self-organization of adequate functional hierarchy by utilizing the constraint of the multiple timescale property and the topological connectivity imposed on the network configuration. These results of the current research could contribute to developments of socially intelligent robots endowed with cognitive communicative competency similar to that of human.

Keywords: Dynamic neural network model, Socially intelligent robot, Self-organization, Compositional semantics, Functional hierarchy

1. Introduction

Recently, studies on socially intelligent robots (Breazeal, 2004; Dautenhahn, 2007) have attracted much attention in the research field of intelligent/cognitive robotics. The main motivation of these studies is to investigate theories and methods for building robots that can perform human-like interactions with other agents, including human as well as other robots, autonomously (Breazeal, 2004). Studies on socially intelligent robots inherit some of their design philosophy, as discussed in behavior-based robotics, sourced from Rodney Brooks (Brooks, 1986) in the late 80s. Although conventional studies on intelligent robots attempted to add behavioral components at a later time, having the basic components on “intelligence” for thinking and cognition built first, the researchers in behavior-based robotics fields did otherwise. By following the thoughts of embodied mind (Rosch, Thompson, & Varela, 1992), they considered that these two processes of thinking and acting should be organized inseparably. Similarly, the researchers in socially intelligent robotics consider the processes of thinking, acting, and communicating as one inseparable process (Billard & Dautenhahn, 1998; Breazeal, 2004; Dautenhahn, 2007).

Recently, not only academia, but also commercial industries, have made great efforts in the development of socially intelligent robots for possible use as home-robots or pet-robots. Such examples can be easily found, including the dog-like robot, AIBO, developed by Sony (Fujita & Kageyama, 1997) and a human-interacting humanoid robot, Pepper, by Aldebaran (Aldebaran, 2014). They proposed new types of home entertainment for family members through interaction with these robots.

Some other researchers, especially in the research field referred to as developmental robotics (Asada et al., 2009; Cangelosi et al., 2010; Lungarella, Metta, Pfeifer, & Sandini, 2003), have tried to apply various psychological aspects evidenced in human infant development, in building cognitive models of socially intelligent robots. At the same time, they attempted to contribute to the understanding of the underlying mechanism or principles for the development of social cognitive functions via their reconstruction in robotics experiments, by utilizing psychologically and neurobiologically plausible models. These social cognitive functions include learning to imitate action demonstrated by others (Billard, 2002; Demiris & Hayes, 2002; Gaussier, Moga, Quoy, & Banquet, 1998; Ito & Tani, 2004; Schaal, 1999), emergence of turn taking skills such as switching between

following and followed among simulated agents (Iizuka & Ikegami, 2004) as well as between human and robots (Nadel, Revel, Andry, & Gaussier, 2004), or joint attention for achieving coordinated behaviors between human and robots (Nagai, Hosoda, Morita, & Asada, 2003; Triesch, Teuscher, Deák, & Carlson, 2006). However, such reconstructions of social cognitive functions through learning have not well addressed the problem of how cognitive competency of systematicity (Cummins, 1996; Fodor & Pylyshyn, 1988) can be developed and also how the contextual flow in particular social cognitive tasks can be captured via iterative learning of the social experiences in the adopted tasks.

The current research aims for reconstruction of cognitive mechanisms with systematicity and context dependency in a neurobiologically plausible manner in a macroscopic sense, which enables robots to perform communicably congruent tasks with human subjects via learning from own sensory-motor experiences. Here, systematicity especially in language processing refers to the cognitive capability of a human to infer the meaning of unknown sentences from known sentences, by extracting compositional semantic rules from them. In the adopted communicably congruent tasks in the current study, we utilize human gestures characterized by systematicity as communicative modality. (It has been shown that human natural gesture recognition capability is also endowed by systematicity (Cassell, Kopp, Tepper, Ferriman, & Striegnitz, 2007; Streeck, 1993).) More specifically, a humanoid robot is tutored to generate specific compositional motor primitive patterns as corresponding to imperative gestures demonstrated by the human subject. An importance here is that the imperative gestures demonstrated by the human subject consists of various combinatorial sequences of movement patterns by following a predefined compositional semantic rules. For example, the robot is requested to respond by generating particular sequences of movement primitives either in order or in reverse order as well as either slowly, normally or quickly as specified by imperative gesture demonstrated.

A technical challenge is to achieve systematicity whereby the robot should become able to infer the underlying meaning or intention for newly demonstrated gesture patterns by means of generalization via learning of prior experienced ones. A difficulty arises because the adopted task uses continuous dynamic patterns of human gesture as well as robot motor response for the communication channel instead of language of discrete symbol sequences. Those dynamic patterns are not always repeatable with variance in their profiles and features. Additionally, in the adopted communicable congruence task, there are no cues that explicitly indicate the structures of the

adopted tasks to the robot. First, there are no explicit cues that indicate types of movement patterns performed by the human subject. The demonstrated movement patterns could be either movement primitives to be memorized for regeneration or commands in specifying order of regular or reverse, or with a speed of slow, normal or fast. The underlying type structures should be learned from scratch out of iterative experience of continuous perceptual patterns in demonstrated gesture. Also, there are no explicit cues to segment on-going task flow between the human demonstration phase and the robot response phase in the course of continuous alternation of these two phases. The turn taking between these two phases should be developed autonomously in the course of learning of examples.

Furthermore, the robot should be able to keep the context of the task flow during each session -- observing the imperative gesture first, and then generating corresponding behaviors, and meanwhile, the accumulated context in the previous session should be reset in the beginning of new session. Such control of the contextual flow, as well as the control of turn taking, are considered to belong to another human specific cognitive capability, which should be developed gradually via iterative learning of task examples. Although the current cognitive task using a particular set of imperative gestures may not replicate our everyday social cognitive behaviors exactly, human, certainly, use gestures to achieve communication characterized by systematicity and context sensitivity (Arbib, 2012; Bowie, 2008; Kendon, 2004). It can be said that the current robotics experiment attempts to model such social cognitive competency of human in an abstract manner.

Here, the technical challenges focused in the current research are summarized as: (1) the adopted MTRNN can achieve generalization by learning in the lower feature perception level by using a limited set of tutoring patterns, (2) acquisition of compositional semantics with generalization for achieving communicable congruence tasks characterized by systematicity, (3) development of cognitive mechanism for controlling the turn-taking process as well as controlling the contextual flow as situated to on-going task processes. The aforementioned technical challenges of targeting development of the cognitive competency characterized by systematicity and context dependency out of lower level perceptual experiences are considered to be novel as compared to those prior-existing studies aiming to reconstruct social cognitive functions via learning, as previously mentioned. Therefore, these technical challenges could contribute significantly to the realization of socially intelligent robots.

For the purpose of accomplishing the aforementioned challenges, the current study takes on

an approach based on the paradigm of dynamical systems and self-organization in modeling the development of target cognitive-behavioral processes, because the research outcomes accumulated for two decades have shown that this approach is one of the best to account for the essence of the embodied cognition (Beer, 2000; Clark, 1999; Kelso, 1997; Lewkowicz & Lickliter, 1995; Port & Van Gelder, 1995; Tani, 1996; Thelen, 1994). The current research especially follows the results from the study conducted by Yamashita and Tani (Yamashita & Tani, 2008), which conjectured on how actional compositionality may be developed by self-organization of particular dynamic neural network models via iterative learning of sensory-motor experiences. They showed that functional hierarchy for generating complex behaviors can be developed through iterative learning of sensory-motor experiences by utilizing the reported multiple timescales recurrent neural network (MTRNN) model. In their study, it was shown that a set of behavior primitives can be learned in the fast timescale dynamics of the lower level subnetwork, whereas sequential combinations of these behavior primitives are learned in the slow timescale dynamics of the higher level subnetwork. The current study is related also to a robotics study on the associative learning between proto-language and behaviors conducted by Sugita and Tani (Sugita & Tani, 2005) by utilizing a version of recurrent neural network (RNN), so-called the recurrent neural network with parametric bias (RNNPB). This study investigated how the compositional semantic rules can be extracted with generalization from the iterated tutoring experience by learning.

The current research attempts to apply these frameworks to realization of a communicably congruent response of robots to humans i.e., robots generating corresponding or expected actions to gestures demonstrated by a human, which is characterized by systematicity and context sensitivity. Detailed analysis on the results of the robotics experiment will clarify how the cognitive competency necessary for achieving the aforementioned communicative skills can be developed in the course of self-organizing adequate dynamic structures in the adopted dynamic neural network model. Next section will describe general idea of the adopted human-robot communicably congruent task which is followed by the descriptions of the employed model and a set of experiments performed.

2. General description of the current task.

Tasks in this paper were designed to investigate how robots implemented by the MTRNN model

becomes able to generate adequate behavioral responses by recognizing compositional communicative gestures demonstrated by human subjects via iterative learning through tutoring. The communicable congruence tasks by robots with human examined in the current study are designed to investigate the technical problems described in the previous section. We designed the following tasks for addressing these technical problems.

Communicably congruent tasks in the current study consist of sequentially-dependent imperative gestures demonstrated by a human and corresponding responses generated by a robot. An imperative gesture is a sequential combination of human movement patterns: 3 different movement primitives, 2 different order commands and 3 different speed commands. Note that the following is a representative task with the most complex experimental conditions. In such condition, one to three human movement primitives were demonstrated in sequence first, followed by an order command and then a speed command. Human movement primitives determine the corresponding robot motor primitives that have to be generated in the response phase by the robot. Order commands are verb-like commands indicating either forward or reverse order in generation of motor primitive sequences. Speed commands are adverb-like commands indicating speed of robot motor pattern, *i.e.* either fast, normal or slow. A corresponding robot response is sequential combination of motor primitives resulting from movement primitives, and its order and speed were determined by order command and speed command. The number of all possible pairs of imperative human gesture and corresponding robot response turns out to be 234. After demonstrating certain human movement patterns, a human subject goes back to the home position and stays for 5-10 time steps. After demonstrating certain robot movement patterns, a robot returns to home position and stops for 12 time steps. Figure 1 shows example pairs of imperative human gesture and robot response in the adopted communicably congruent tasks. (We also conducted simpler tasks for other experiment purpose as described later.)

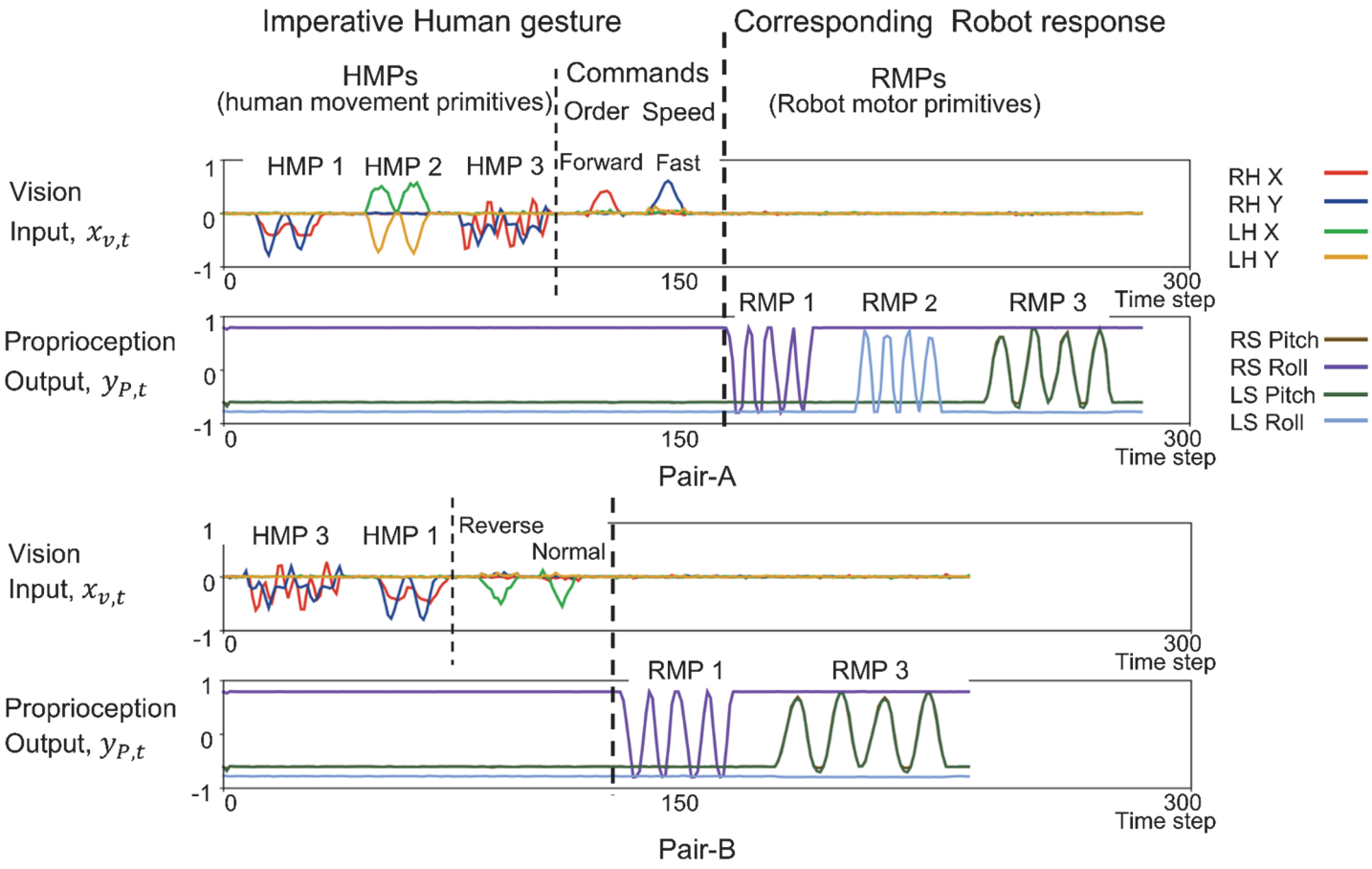

Figure 1. Examples of pairs in the communicably congruent task. A human demonstrates an imperative gesture pattern, and after some delay, a robot generates the corresponding response. An imperative gesture consists of the movement primitives, order and speed commands.

It is noted that the robot has to learn how to produce a communicably congruent response via iterative supervised training by human tutors. For each tutoring session to produce a communicably congruent response, the robot receives continuous sequences of movement patterns observed in the human gesture demonstration, followed by continuous sequences of patterns of its own movements in terms of proprioception (motor joint angles) provided through guidance by the human tutors. The session of the tutoring is repeated for many different pairings of the imperative human gesture patterns and their corresponding robot responses. The MTRNN model described later is used for batch learning of the visuo-proprioceptive sequence patterns collected through the accumulated tutoring sessions.

The aforementioned communicably congruent tasks are not trivial for the following reasons. First, both imperative human gestures and the corresponding robot responses are a combination of movement patterns. Second, the imperative gesture is generated by following compositional semantic

rule; a human subject can demonstrate one to three movement primitives before demonstrating commands. Third, the same movement patterns could have different meanings depending on their previous pattern's type. For example, in Pair-B in Figure 1 although two succeeding reverse order command and the normal speed command are in the same shape, their types are different. Note that there is no explicit cue, *e.g.* symbol or label, in sensory-motor flows that network can utilize to segment between the human movement primitives and the commands or the human gesture phase and the robot response phase. The proposed network has to develop cognitive function to segment patterns and phases and to extract the underlying meanings and rules among those segments by just going through iterative learning of the continuous sensory-motor flows.

In some experiments described in the current paper, test trial was conducted for a single session consisting of a pair of human gesture part and robot action generation part. In other experiments, test was conducted for multiple concatenated sessions without break between the sessions. The latter case is more challenging because the end of each session has to be recognized autonomously by the robot.

Following are the details of the human movement patterns and robot motor primitives. The human movement patterns are left and right arm movements. Trajectory data of the arm movement patterns were recorded by tracking xy position of the palm positions through the KINECT skeletal tracking system (Zhang, 2012) while a human subject performed the movement patterns. For all dimensions, the position value for initial position was 0, and the range of the position values was [-1:1] for Experiment 1 and [-0.8:-0.8] for Experiment 2 and 3. We employed the humanoid robot NAO as an intelligent robot agent for the communicably congruent task. To get trajectory data of the motor primitives of the robot action, the human experimenter guided both arms of the NAO by moving NAO's arms. While guiding, we encoded 4 joint angles comprising shoulder rolls and pitch angles of both arms. Ranges of the angles were [-0.8:0.8]. The sampling rate of both human movements and robot motor primitives were 10Hz.

Human movement patterns and robot motor primitives adopted in this study are described in Figure 2. Figure 2 (a) shows correspondences between the human movement primitive patterns demonstrated and the robot motor primitive patterns responded. There are three pairs. Figure 2 (b) shows the forward and reverse order command patterns demonstrated by the human subject. Figure 2 (c) shows the speed command patterns for slow, normal and fast. It is noted that the normal speed

command case is tricky. The rule here is that when the same pattern with the order command demonstrated immediately before appears, it should be interpreted as the normal speed command. This type of communicative semantic rule was adopted for the purpose of examining the capability of the model to cope with context dependent parsing of demonstrated patterns.

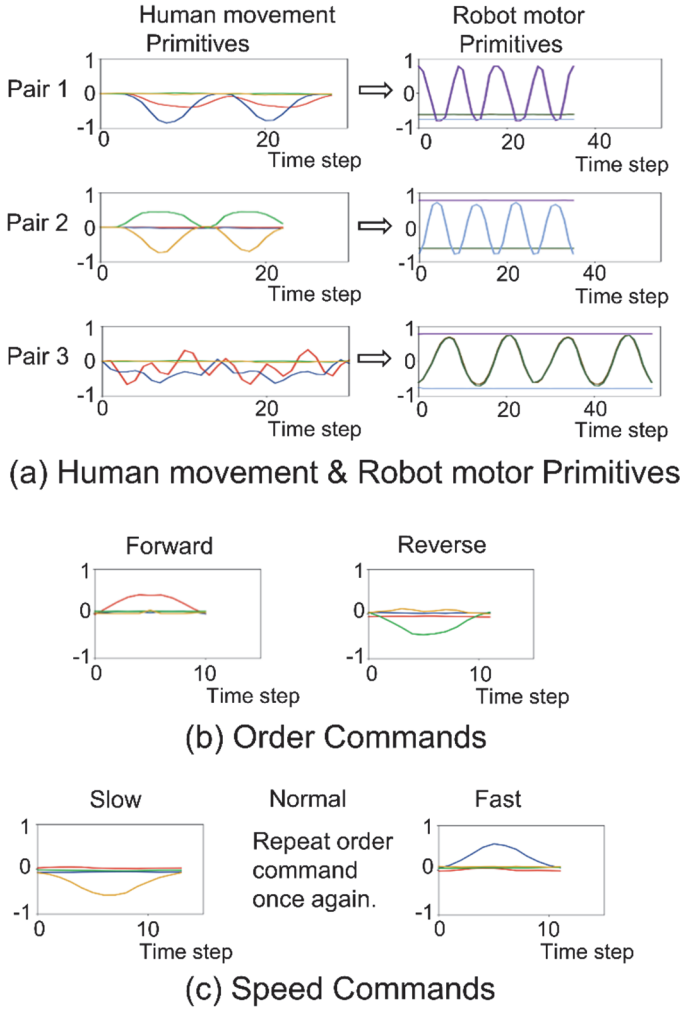

Figure 2. Human imperative gesture movement patterns and robot motor primitive patterns used in the task. (a) Three corresponding pairs between human movement primitives and robot motor primitives, (b) order command patterns for forward and reverse and (c) speed command patterns for slow, normal and fast.

3. MTRNN used in the current experiment

The MTRNN model used in the current study consists of multiple subnetworks characterized with different timescale dynamics in neural activity. The network is expected to generate adequate motor response in adequate timing by receiving continuous flow of visual perception from Kinect without

being provided with task related cues by the experimenter. In the previous study for compositional action generation using MTRNN (Yamashita & Tani, 2008) or (Tani, 2003), the intention to select which action sequence to be generated was given by the experimenter by setting specific values to the initial states of the internal (context) units of the MTRNN or to the PB unit states of the RNNPB. In the current study, the intention for generating a particular motor response is not given by the experimenter but expected to be built up in the internal (context) dynamics while perceiving continuous flow of the human gesture patterns by setting the initial state of the internal dynamics with neutral value. After the human ceases demonstration of the gesture, the motor response is autonomously generated as reflected with the top-down intention built up in the internal dynamics so far. It is highly speculated that such mechanism of turn taking from observation of the gesture to generation of own action can be developed in the dynamical structure of the network model if the network is trained with sufficient amount of tutoring for pairing observation of the gesture and generation of own corresponding motor response. Next subsections describe details of the adopted network model.

3.1 Model architecture.

We extended the MTRNN model (Yamashita & Tani, 2008) which is composed of multiple levels of subnetworks characterized by different timescale dynamics. Each subnetwork contains a set of leaky-integrator type neural units of which dynamic follow the following differential equation Eq. (1).

$$\tau_i \dot{u}_{i,t} = -u_{i,t} + \sum_j w_{ij} x_{j,t} + \sum_l w_{jl} c_{l,t-1} \quad (i, l \in C \wedge j \in I) \tag{1}$$

where $x_{j,t}$, $u_{i,t}$, $c_{l,t}$ represent the jth input, ith internal state, and lth context activation value at time t, respectively, τ_i is the time constant of the ith input-output and context units, C, I and O are the neuron indices of the context, input, and output layers, and w_{ji} is the ith unit to jth unit connectivity. Here, context units are neural units which do not have direct inputs and outputs. From Eq. (1), it can be seen that the larger the time constant, the longer the neural activation state integrated through time. Yamashita and Tani (Yamashita & Tani, 2008) showed that a group of neural units with larger or smaller time constant tends to capture longer or shorter time correlation structures through learning, respectively.

A typical architecture of MTRNN adopted in the current experiments is shown in Figure 3.

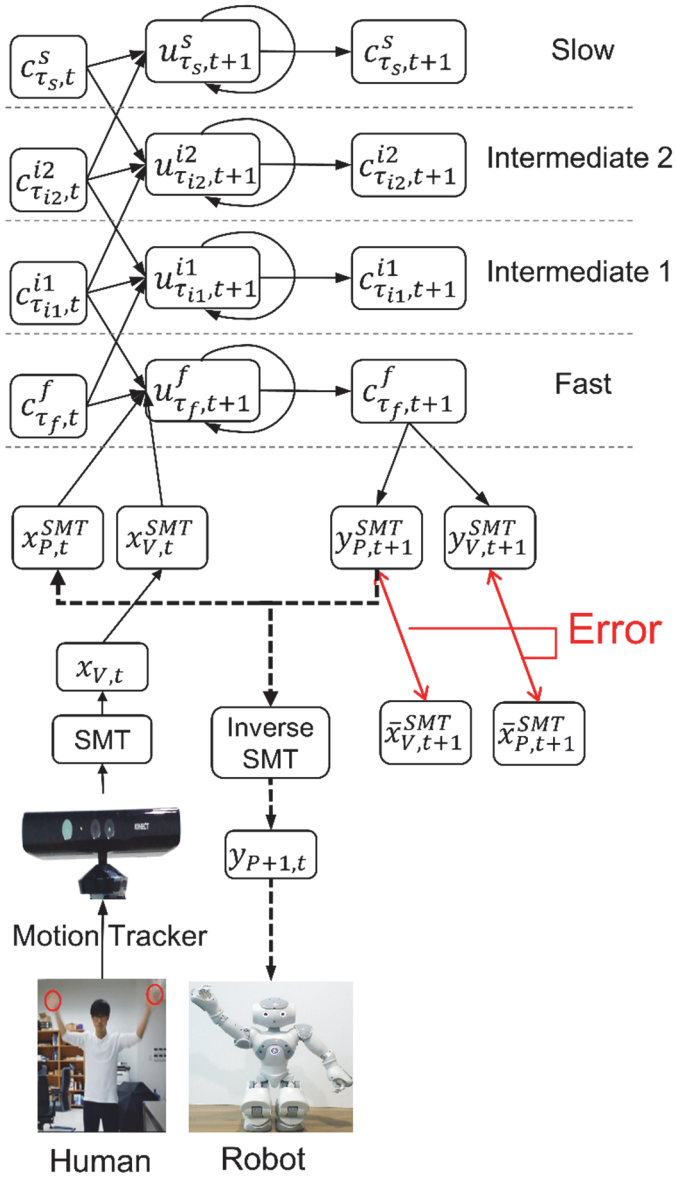

Figure 3. The whole system architecture for the performing the communicably congruent tasks with MTRNN. SMT denotes softmax transformation. $x_{V,t}^{SMT}$ and $x_{P,t}^{SMT}$ are the vision and proprioception inputs, respectively. $x_{V,t+1}^{SMT}$ and $x_{P,t+1}^{SMT}$, are vision and proprioception prediction outputs, and $\bar{x}_{V,t+1}^{SMT}$ and $\bar{x}_{P,t+1}^{SMT}$ are their targets. τ is the time constant of the context unit. c_t and u_t are the activation value and internal state of the context unit at time step t, respectively.

This particular network architecture consists of four subnetworks, the fast context dynamics subnetwork assigned with the smallest time constant τ_1, the intermediate-1 context dynamics subnetwork with next small τ_2, the intermediate-2 context dynamics subnetwork with medium τ_3, and the slow context dynamics subnetwork with the largest τ_4, in which the input and output units are connected to the fast context dynamics subnetwork. It is, however, noted that number of subnetworks

could be different depending on the task property, as seen in the later described analysis.

The input is sent from the motion tracker (4 dimensional position data measured for left hand and right hand of the human subject as denoted by $x_{V,t}$) and from the NAO robot guided by the experimenter (the encoder reading of 4 DOF joints as denoted by $x_{P,t}$). We used the softmax transformation to the inputs. Softmax transformation is independently applied to each dimension of the Kinect inputs and the proprioception inputs. These transformed data were used as the input ($x_{V,t}^{SMT}$ and $x_{P,t}^{SMT}$) and target-output ($\bar{x}_{V,t+1}^{SMT}$ and $\bar{x}_{P,t+1}^{SMT}$) of the network. The following equation, Eq. (2), describes the softmax transformation (Bishop, 2006):

$$p_{ij,t} = \frac{exp\frac{-\|k_{ij} - k_{i,t}^{sample}\|^2}{\sigma}}{\sum_{j \in Z} exp\frac{-\|k_{ij} - k_{i,t}^{sample}\|^2}{\sigma}} \tag{2}$$

where $k_{i,t}^{sample}$ indicates position value of the ith dimension at time t, k_{ij} is the value of the jth element of the ith dimension's reference vector, σ is a constant that determines sharpness of the distribution, and p are transformed vectors. Values of the elements of the reference vectors are calculated by the following Eq. (3):

$$k_{ij} = -B_i + 2\frac{B_i}{l(i) - 1}(j - 1) \tag{3}$$

where $l(i)$ is the length of the ith dimension's reference vector and B_i is the ith dimension's upper bound. We use 0.01 as a σ, and 9 for $l(i)$ for all i.

A part of output values is sent to the robot controller as 4 DOF motor joints target denoted by $y_{P,t+1}^{SMT}$ and the remained part provides next time step prediction of the 4 dimensional position data of the human subject hands. The adjacent context dynamics subnetworks are connected mutually by synaptic weights but not for distant subnetworks. The detail configuration of the architecture such as number of the subnetworks can vary depending on the purpose of each experiment.

3.2 Forward dynamics and training

For given connectivity weights, the forward dynamics of all neural units including output units and context units in different timescale subnetworks are computed by following Eqs. (4~6).

$$u_{i,t+1} = \begin{cases} \left(1 - \frac{1}{\tau_i}\right) u_{i,t} + \frac{1}{\tau_i}\left(\sum_j w_{ij} x_{j,t} + \sum_l w_{il} c_{l,t} + b_i\right) & (i, l \in C \wedge j \in I) \\ \sum_l w_{il} c_{l,t+1} + b_i & (i \in O \wedge l \in C) \end{cases} \quad (4)$$

$$c_{i,t+1} = f\left(u_{i,t+1}\right) \quad (5)$$

$$y_{ij,t+1} = \frac{\exp\left(u_{ij,t+1}\right)}{\sum_k \exp\left(u_{ik,t+1}\right)} \; (i \in V, P \wedge j, k = 1,2, \ldots, l(i)) \quad (6)$$

where V and P are sets of indices corresponding to vision and proprioception, and *f()* is hyperbolic tangent. The output $y_{ij,t+1}$ is computed by using the softmax function. This softmax output activation function is used to help learning by making the output activate sparsely. The forward dynamics generates a sequence of one-step prediction for the output vector. In the current study, the initial internal state of each context unit is set with a neutral value of 0.0.

The training of the network is conducted by means of off-line supervised teaching with target training sequences. The training sequence data consists of 4 dimensional motion tracker values at each step obtained from Kinect and 4 dimensional motor joint positions at each step obtained from the robot guided by the experimenter. Each sequence data is obtained from each trial of tutoring for one session (a pair of imperative gesture and corresponding motor generation) or across multiple sessions iterated depending on the experiment task. Multiple training sequences are used to train the network such that generalization can be achieved in learning.

A learning scheme referred to as back-propagation through time (BPTT) algorithm (RUMELHART, Hinton, & Williams, 1986) is utilized for training of the network. The network was trained to optimize its learnable parameters to minimize the following objective function. The objective function is the Kullbak-Leibler divergence between the target in the next step and its prediction output, as described in the following Eq. (7).

$$E = \sum_t \sum_{i \in O} y_{i,t+1} \log\left(\frac{\bar{x}_{i,t+1}}{y_{i,t+1}}\right) \quad (7)$$

where $\bar{x}_{i,t+1}$ is next step target output. The learnable parameters consist of connectivity and bias are updated in a direction of minimizing prediction error, *i.e.* opposite direction to that of the gradient $\frac{\partial E}{\partial \theta}$ as described in Eq. (8).

$$\theta(n+1) = \theta(n) - \alpha \frac{\partial E}{\partial \theta} \quad (8)$$

where θ indicates learnable network parameters consist of connectivity and bias, s is the index of training sequences, α is learning rate. $\frac{\partial E}{\partial w_{ij}}$ and $\frac{\partial E}{\partial b_i}$ are computed by following Eqs. (9) and (10).

$$\frac{\partial E}{\partial w_{ij}} = \begin{cases} \frac{1}{\tau_i}\sum_s\sum_t \frac{\partial E}{\partial u_{i,t}} x_{j,t} & (i \in C \wedge j \in I) \\ \frac{1}{\tau_i}\sum_s\sum_t \frac{\partial E}{\partial u_{j,t}} c_{j,t-1} & (i \in C, O \wedge j \in C) \end{cases} \tag{9}$$

$$\frac{\partial E}{\partial b_i} = \frac{1}{\tau_i}\sum_s\sum_t \frac{\partial E}{\partial u_{i,t}} \tag{10}$$

The delta error of the ith unit at time t is calculated from the following Eq. (11).

$$\frac{\partial E}{\partial u_{i,t+1}} = \begin{cases} y_{i,t+1} - \bar{x}_{i,t+1} & (i \in O) \\ \sum_j \frac{\partial E}{\partial u_{j,t+2}} \left[\delta_{ij}\left(1 - \frac{1}{\tau_i}\right) + \frac{1}{\tau_j} w_{ji} f'(u_{i,t+1})\right] & \\ \quad + \sum_k \frac{\partial E_t}{\partial u_{k,t+1}} \left[w_{ki} f'(u_{i,t+1})\right] & (i,j \in C \wedge k \in O) \end{cases} \tag{11}$$

where, $f'()$ is the derivative of the hyperbolic tangent, δ is Kronecker delta function, and E_t is the objective function at time step t. For the ith context unit, the delta error is recursively calculated as described in second line of Eq. (11). In this recursive calculation process, τ_i determines how much the delta error of ith context unit will be preserved with slow attenuation, and it will filter out instant fluctuations in the backpropagation process. Therefore, for example, the slow dynamics subnetwork that has a large time constant can learn long timescale correlation in the input and target. On the contrary, fast dynamics subnetwork can learn short timescale correlation independent from the long timescale correlation without filtering out fast changes in delta error and fast attenuation of it.

As can be seen in Figure 3, the MTRNN is a type of deep neural network, which has a deep hierarchical structure of multiple subnetworks and each of them has different timescale dynamics with a different time constant. Therefore, the network would learn multiple timescale correlations and self-organize into a functional hierarchy between multiple subnetworks by utilizing its hierarchical structure and multiple timescale dynamics. As a result of the learning by means of prediction error minimization, the network becomes able to represent the states of both the environment and itself, and makes links between them at the level of the slow dynamics subnetwork. This permits the network to develop adequate cognition for communicating with others in the outer world.

The initial values of weights and biases are randomly set with a Gaussian distribution. The range of the Gaussian distributions are [-0.1, 0.1] for the weights and [-1, 1] for the biases. The initial

learning rate was set as $^{0.1}/_{Ttotal * d}$, where $Ttotal$ means summation of the time step over all training sequences, and d is dimensionality of the output.

In order to accelerate the learning speed as well as to achieve better generalization capability, a scheme of adaptive learning rate was employed. Duffner and Garcia (Duffner & Garcia, 2007) introduced the adaptive learning rate method in which learning rate is adjusted based on the validation errors, and they showed that the method can improve both convergence speed and generalization capacity. The training algorithm employed in the current model employs this scheme. Details of the adaptive learning rate algorithm is given in Appendix A. Furthermore, we noticed that it is technically important to select the best learning parameter values of the weights and biases developed in the course of the training process because the learning error often fluctuates as the training proceeds in the preliminary study. We employed a scheme of selecting the best learning parameters obtained in the course of the entire training process, whose scheme is described in Appendix-B.

4. Experiments and analysis of the results

A series of experiments on human-robot communicative skills based on learning was conducted for the purpose of examining a set of essential problems described previously. More specifically, the Experiment-1 examines the capability of the network model for the feature level generalization in learning of pairs of relatively simple gestures and motor responses. Experiment-2 investigates the generalization capability of the network model at the cognitive level in extracting compositional semantic rules from demonstrated imperative gestures. Experiment-3 examines the development of the cognitive mechanism, by which adaptive control of the contextual memory could be achieved.

4.1 Experiment 1: Lower feature level generalization

Human gesture movement patterns cannot be generated steadily. Even when the human subjects are asked to repeat the same movement patterns, features of patterns such as speed or amplitude can be vary at every trial. However, it is assumed that human can recognize the same category of movement pattern regardless of variations in various features by means of generalization by extracting invariant features. Then, a question is that if the MTRNN model can exhibit a similar robust recognition for

demonstrated movement patterns by means generalization through learning, how much amount of variant patterns in combination of different features are necessary as for training exemplar for achieving successful generalization.

In order to investigate the aforementioned problem, an experiment was conducted by adopting a simpler task setting as compared to the general task described in the previous section. An imperative gesture in this experiment consists of one movement primitive selected among a total of 3 different primitives, followed by the forward order and the normal speed commands. The movement primitive patterns were demonstrated with variations of 7 different amplitudes (AMP) and 7 different velocities (VEL). Then, the robot had to generate the corresponding motor primitive pattern with the normal amplitude and with the normal speed. Movement primitives composed of different variant features, in these experiments, were made in the following manner. First, each primitive was demonstrated in 7 different trials and within each trial it was introduced in differing amplitudes, 70, 80, 90, 100, 110, 130 and 150%, respectively. After that, the primitives were artificially stretched in time. The mean amplitudes of each movement primitive were 0.288, 0.364, and 0.514, and the standard deviations of amplitudes of each primitive were 0.070, 0.092, and 0.128, respectively. The performance of generalization in learning was tested with different networks trained with 4 different size of training data sets. Table 1 describes variations of each movement primitive pattern prepared for the training data set.

Table 1

VEL / AMP	50	70	80	100	110	130	150
70							
80							
90							
100				P1[a] P2[b] P3[c]			
110							
130							
150							

(a) Training data set 1

VEL / AMP	50	70	80	100	110	130	150
70							
80				P2			
90							
100		P1		P1 P2 P3		P1	
110							
130				P2			
150							

(b) Training data set 2

VEL / AMP	50	70	80	100	110	130	150
70							
80		P1		P1 P2		P1	
90							
100		P1 P3		P1 P2 P3		P1 P3	
110							
130		P1		P1 P2		P1	
150							

(c) Training data set 3

VEL / AMP	50	70	80	100	110	130	150
70							
80		P1 P2 P3		P1 P2 P3		P1 P2 P3	
90							
100		P1 P2 P3		P1 P2 P3		P1 P2 P3	
110							
130		P1 P2 P3		P1 P2 P3		P1 P2 P3	
150							

(d) Training data set 4

[a]P1 = human movement primitive 1; [b]P2 = human movement primitive 2; [c]P3 = human movement primitive 3

The configuration of each training data set.

The human movement patterns were generated with seven different amplitudes: 70, 80, 90, 100, 110, 130, and 150%, and seven different velocities: 50, 70, 80, 100, 110, 130 and 150%. Consequently, each human movement primitive pattern can be demonstrated with 49 different variations using combinations of the amplitude feature variation and the velocity one. Human movement primitives 1, 2, and 3 were denoted by P1, P2, and P3 in the tables, respectively. Each slot represents different combination of velocity variation and speed variation. For example, Table 1 (a) denotes that P1, P2 and P3 are prepared with only 100% amplitude and 100% velocity cases in the training dataset 1. Table 1(b) denotes that more variations are added for P1 velocity and P2 amplitude in the training dataset 2. Table 1(c) and Table 1(d) denote that further more variations are added in both velocity and amplitude features for all movement primitive patterns. Three sample sequences used for each combination of amplitude and velocity for obtaining enough sequences to use the training method utilizing validation data set (see Appendix-A, B). The ratios of training pairs to test pairs in 4 different data set were 3/144, 7/140, 15/132, and 27/120.

For the test of the generalization performance of the trained network, all imperative gesture movement primitive patterns which were not utilized in the training dataset were used for and success rate of responding with correct motor primitive patterns was counted. The number of context dynamics subnetworks was set with 3. The first context dynamics subnetwork consists of 30 units with tau = 2, the intermediate one does of 20 units with tau = 5, the slow context dynamics subnetwork does of 6 units with tau = 30. The network was trained with 4 different datasets denoted in Table 1 with the training method explained in Section 3.2. For each data set, the network was trained three times with initial parameters differing each time. The number of training iteration was with 300,000. The best epoch in the training result was selected by the prescribed criteria (see appendix-B). After training with each dataset, the success rate in the congruent response by the robot was obtained as follows. The success rate was measured by counting the number of the tasks that the network successfully performed. If the mean KL-divergence between the proprioception target sequence and the output sequence was lower than 0.01 during the motor response, the trial was considered as successful.

The results for all training datasets are summarized in Figure 4. It can be seen that the success rate for unlearned test trial case increases as the number of the training sequences is increased. In the cases of using the dataset 3 and the dataset 4, the average success rates for the test trials becomes more than 92%. It can be said that the network can achieve relatively good generalization in recognition of demonstrated gesture patterns with the feature level variation by learning with limited amount of training dataset.

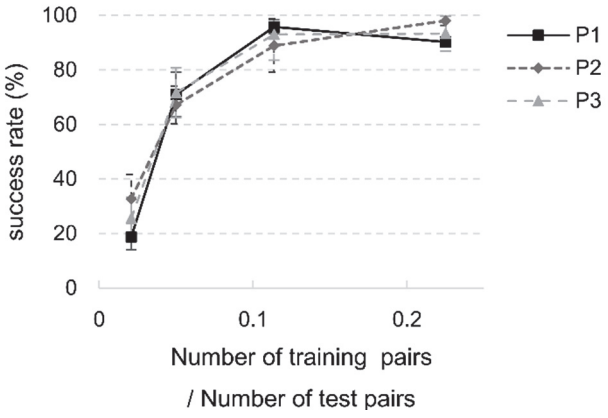

Figure 4. Comparison of the success rates of test data for 4 different sets. P1, P2, and P3 denote movement primitives in gestures. For each data set, the success rates of test pairs containing different movement primitives, P1, P2, and P3, are individually shown in this figure. Error bar indicates the degree of standard deviation.

4.2 Experiment 2: Development of systematicity

Acquiring compositional semantic rules from gesture-response pairs by learning is an essential problem for development of human-like socially communicative robots. This is because humans can generate numerous expressions even by combining a small number of components in different order. For example in the current experiment, the human subject can demonstrate 234 different imperative gestures which can be generated by arranging only 3 different human primitives in sequential combinations, 2 order commands, and 3 speed commands in different sequences. Furthermore, same movement primitive patterns in human gestures can have different meanings depending on the context, such as the case of the semantic rule adopted for the speed command as described previously. Because it is tedious to train all possible combinations of gesture patterns to the robot, it is highly desired that the network model can learn to extract the underlying compositional semantic rules from partial training exemplar.

Sugita and Tani (Sugita & Tani, 2005) showed that semantic compositionality comprised of combinations of 3 transitive verbs and 3 object nouns which is mapped to 9 different action categories can be learned with generalization by using 7 pairs out of 9 all possible combinations for the training. The current experiment examines the case of learning more complex communicably congruent tasks where the number of all possible combinations in imperative gesture is 234.

In Experiment 2-1, we examine the capabilities of the MTRNN by training with a part of the all possible imperative human gesture and corresponding robot response pairs, and tested by using the unlearned pairs. The task design follows the one described in Section 2 in which exact profiles of the human movement primitive patterns and the robot motor primitives patterns used in this experiment are shown in Figure 2. We examine the effect of amount of training data on the performance of generalization in learning by repeating the learning and test experiment with changing the training dataset. In Experiment 2-2, we investigate the contribution of the multiple timescale property assigned to the subnetworks as well as topological organization among the subnetworks to the task performance by changing the timescale as well as the topological connectivity condition.

4.2.1 Experiment 2-1: Acquisition of compositional semantic rules

We conducted the experiments of learning and testing by changing the amount of training pairs, the

one with 156 pairs and the other with 78 pairs. In training condition 1, two thirds of all possible pairs, *i.e.* 156 pairs, were used for training dataset and the remaining 78 were used for test dataset. The training data set included the validation dataset (see the appendix-A) which were composed of 12 pairs. In training condition 2, 78 pairs were used for training data set including 6 pairs used for the validation dataset, and the remaining 156 pairs were used for the test data set. The number of context dynamics subnetworks was set with 4. The fast context dynamics subnetwork consists of 30 units with tau= 2, the intermediate-1 one does of 30 units with tau= 5, the intermediate-2 one does of 20 units with tau= 10, the slow context dynamics subnetwork does of 20 units with tau= 60. The network was trained 3 times with different initial learnable parameters.

The experiment resulted as follows. In training condition 1, the average success rate was 87.0% for the test sequences and 99.4% for the training sequences. In training condition 2, the network showed poorer result with average success rate of 46.3% for the test sequences and 93.3% for the trained sequences. The network showed better generalization result when amount of training pairs was increased as expected. (A demonstration video of an example of the robot performance in one of the adopted tasks, after trained in training condition 1, can be seen by clicking here or in http://neurorobot.kaist.ac.kr/video/Gibeom_NN_demo_video_20150213.mp4).

Figure 5 and Figure 6 illustrate how the slow and fast context dynamics developed, in test trials, respectively, along with perception of the human gesture patterns and generation of the motor response after the network was trained in training condition 1. Figure 5 shows activities of representative slow context units along with the vision (Kinect) inputs and the motor (proprioception) outputs in different test trials.

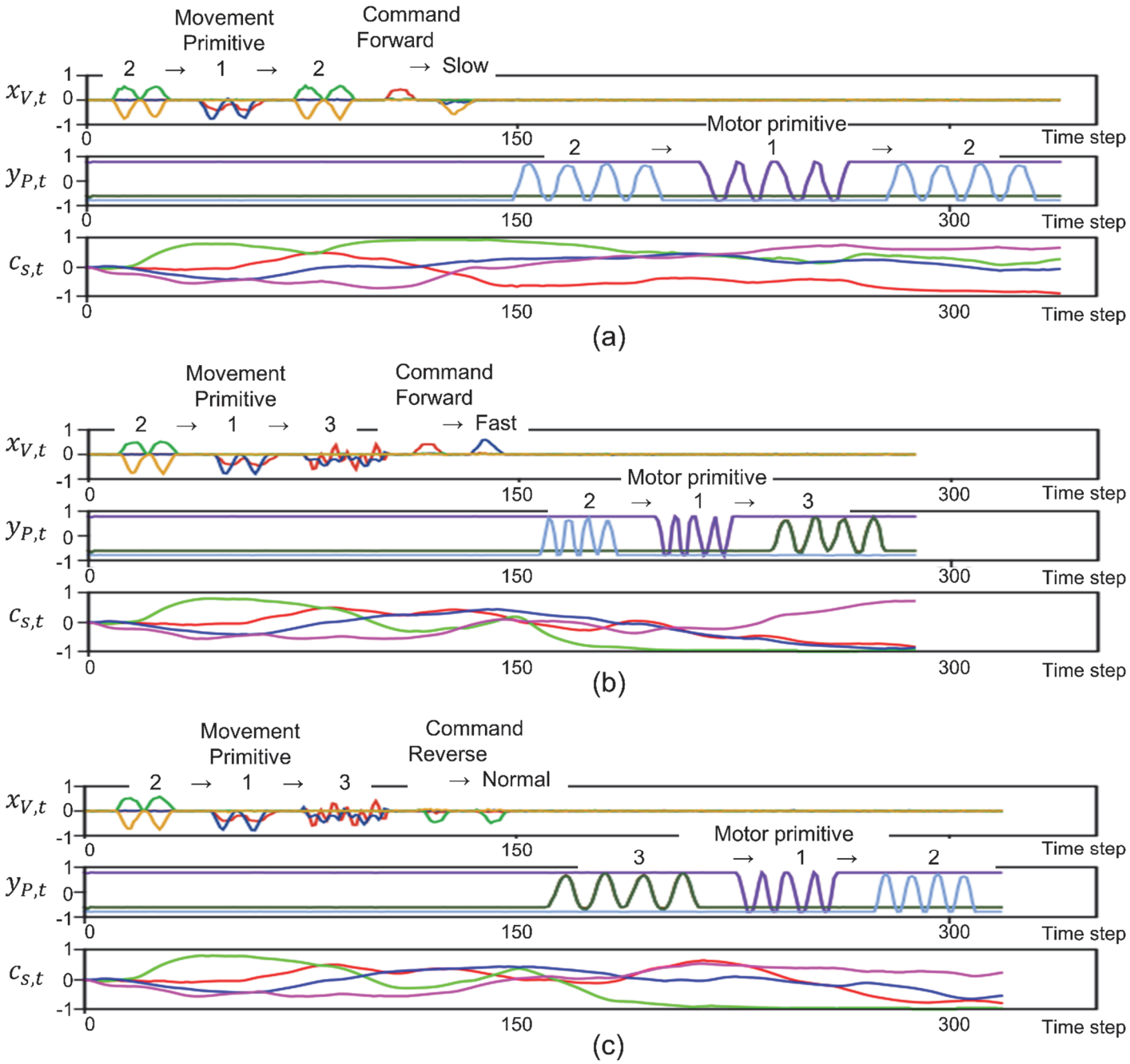

Figure 5. Three examples of test trial results in (a), (b) and (c), plotted with the slow context units activity. $x_{v,t}$, $y_{p,t}$, and $c_{s,t}$ indicate the vision (Kinect) inputs, the motor (proprioception) output and the slow context units activity.

Figure 5 (a) and Figure 5 (b) shows a situation of two test trials that the same movement primitive sequence was demonstrated until the 2nd primitive and different primitives were demonstrated for the third primitive. It can be seen that the slow context unit activities in these two trials were the same before the third movement primitive was demonstrated and they developed differently after different movement primitives were demonstrated for the third primitive. The similar observation can be made

by comparing Figure 5 (b) and Figure 5 (c) where the slow context activities became different after different order commands were demonstrated. This implies that the slow context activity represents the current context by integrating the gesture sequence patterns perceived in the past. Therefore different context activities are developed at the end of the demonstration phase which represent the intention of the human subject expressed by combination of imperative gesture patterns. This integrated activity in the slow context units triggers generation of the corresponding motor primitive sequences.

Figure 6 illustrates activity of the representative fast context units, along with the vision (Kinect) inputs and the motor (proprioception) outputs in different test trials.

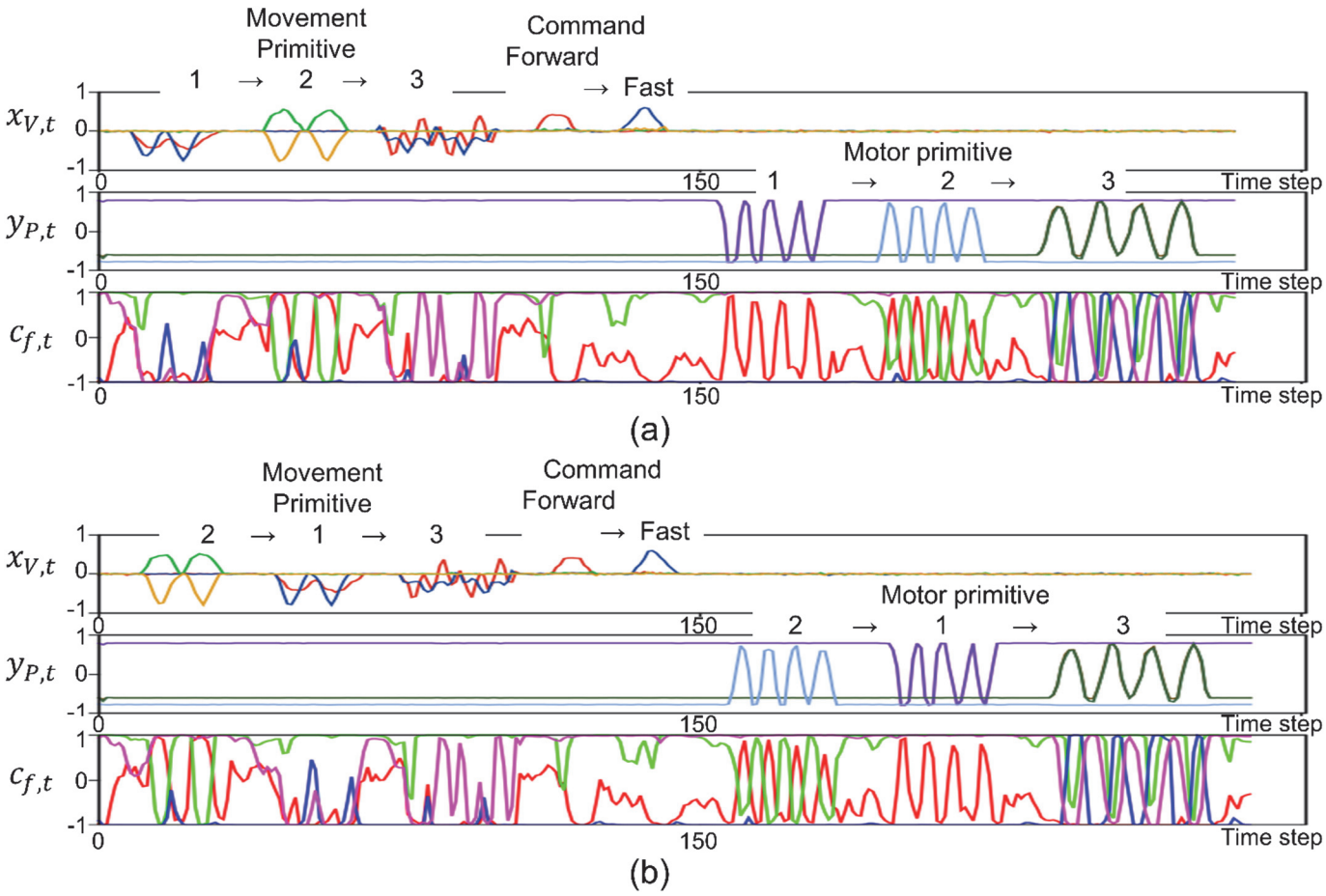

Figure 6. Two examples of the test trial results in (a) and (b), plotted with the fast context activities. $x_{v,t}$, $y_{p,t}$, and $c_{f,t}$ indicate the vision (Kinect) inputs, the motor (proprioception) outputs and the fast context units activity.

It can be seen that the fast context unit dynamics develops quite differently from the slow one by seeing two trial cases shown in this figure. It can be seen that the profiles of the fast context unit activity become quite similar when perceiving the same movement primitives or generating the same

motor primitives in these trials. It can be said that the fast context unit activity develops in terms of one to one mapping from the current on-going perception or the motor generation.

To provide more statistical evidence in a quantitative manner, the overall behavior of the network was analyzed as follows. For each of the training conditions 1 and 2, variances in the changes of the internal states of the context units in both fast and slow dynamics subnetworks in the Case-A, after encountering all slightly different movement primitive patterns belonging to the same category of movement primitive and in the Case-B, after encountering all completely different movement primitive patterns belonging to different category of movement primitive are computed. After that, the rates of the variances on these two cases (the variance in the context activation changes in the Case-B divided by the one in the Case-A) are computed for fast and slow dynamics subnetworks in each training condition. For both training conditions 1 and 2, the average rates measured in slow context dynamics subnetwork were 18.25 and 25.71, respectively. For both training conditions, the rates were measured in fast context dynamics subnetwork were 1.10 and 1.08. For both training conditions, the context units in the slow dynamics subnetwork showed much larger variance upon the introduction of different category of movement patterns compared to encountering the same category of movement patterns. In contrast to the slow context units, the fast context units showed quite small variance. This result is analogous to the aforementioned interpretation that the slow dynamics subnetwork represents the current context by integrating patterns perceived for long period, fast one develops in terms of one-to-one mapping from the current on-going perception, or the motor generation.

To the sum, it can be said that the slow dynamics subnetwork is successful in extracting longer time correlated structures such as semantic rules from observed perceptual sequences. The first dynamics subnetwork, on the other hand, involves with processing detail features of on-going perceptual or motor primitive patterns. These observations are analogous to the previous experiment results reported by Yamashita and Tani (Yamashita & Tani, 2008) which indicated that MTRNN can develop functional hierarchy for generating complex motor behaviors in which the fast dynamics subnetwork encodes each motor primitive patterns and the slow dynamics subnetwork does for concatenation of them. The current experiment results, however, show furthermore that the adopted MTRNN model can recognize the intention of the human partner by extracting complex semantic rules latent in the stream of demonstrated patterns.

4.2.2 Experiment 2-2: Importance of the topological connectivity among subnetworks and multiple timescales property

The results of Experiment 2-1 implied that subnetworks assigned with different timescale dynamics develop different functions necessary to achieve the adopted task. The current subsection explores how configuration of the network including the topological connectivity among subnetworks and the multiple timescale structures assigned to subnetworks can affect the task performance. For this purpose, we conducted the following set of the learning experiment by employing the same task setting shown in Experiment 2-1.

The first set of experiments examined the condition of allocating four different subnetworks consisting of alternating the time constants and number of context units of the slow context dynamics subnetwork. Ten different network settings were tested. In the case of the first five settings, five different time constants were assigned to the slow context units, 20, 40, 60, 80 and 100, while the time constants of other context units were fixed as the same time constants used in Experiment 2-1. As the time constant of the slow context unit increased, the time constant differences between the slow dynamics subnetwork and other subnetworks increased. This difference is interpreted as the multiple time scale property is enhanced. In the case of the other five settings, five different amounts of slow context units were assigned: 20, 40, 60, 80 and 100. For both cases, the remaining network hyper-parameters were the same as the network used in Experiment 2-1. The second set of experiments examined the condition of allocating only a single subnetwork with alternating the time constants and number of slow context units. It should be noted that the second set of experiments were conducted under the same condition as the first set, with difference only in the topological connectivity of the network. All neural units assigned with different time constants were fully connected in a single subnetwork. The results of the task performances for untrained test cases after the network being trained on the aforementioned different network settings are summarized in Figure 7.

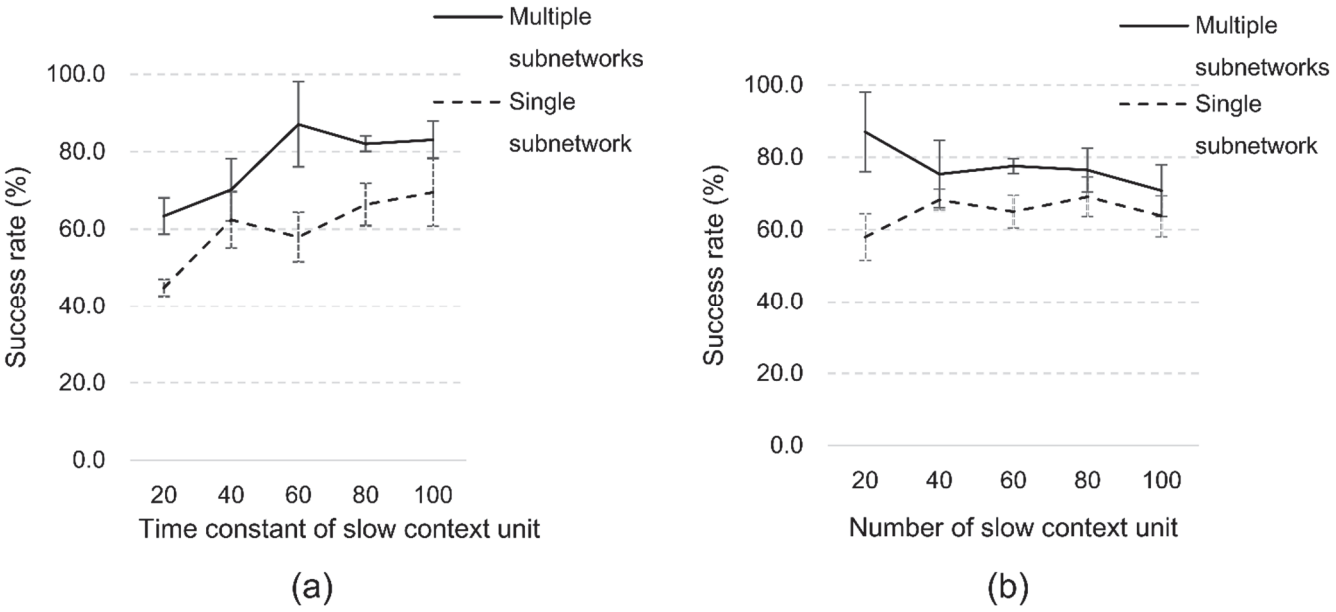

Figure 7. Success rates for test trials with the network on different network configurations of multiple subnetworks and single one: a) with different time constants of the slow contexts and b) with different numbers of the slow context condition. Error bar indicates the degree of standard deviation.

Figure 7 demonstrates that the network consisting of multiple subnetworks outperforms the network consisting of only a single subnetwork. As can be seen in Figure 7 (a), for both networks consisting of single and multiple subnetworks, the networks with large time constants of the slow context units as compared to other context units and with a distinctive multiple timescale property performed better than others. Although the networks showed sensitivity to the parameters, such as to the number of slow context units that can be related to the overfitting problem, as shown in Figure 7 (b), the best performance for the generalization test was obtained in the case of adopting multiple subnetworks configuration with distinctive multiple time scales property among the subnetworks. These results imply that both topological connectivity among subnetworks and multiple timescales property are essential in development of required functional hierarchy with generalization. These observations are analogous to the prior research results that multiple timescale property adequately set in multiple subnetworks can result in better organization of functional hierarchy (Yamashita & Tani, 2008) and also that the bottleneck connectivity between the higher level and the lower one is beneficial in development of adequate functional segregation between different levels (Paine & Tani, 2005). (Experimental results that compared the performance of the 4 different network configurations, (1) multiple subnetworks and multiple time constants, (2) multiple subnetworks with a single time constant, (3) single subnetwork with multiple time constants, and (4) single subnetwork with a single

time constant, can be found in (Park & Tani, 2015))

4.3 Experiment 3: Development of cognitive capability of controlling the contextual memory

In the previous experiment, each session was performed independently as terminated after the robot response phase. However, it might be more natural if the session iterates continuously without termination. This, however, brings another technical challenge that how the robot can segment task sequence into each session without receiving explicit cues of indicating onset or end of the session. In the iterated session case, the network should be able to keep the activity of context units to develop while receiving the sequences of gesture movement primitives. When the gesture demonstration is finished, the network should be able to trigger the turn taking for generating own motor response. When the motor response is finished, the context states should be reset for preparing for next session to be started. This sort of task process requires the network to acquire a certain cognitive mechanism for controlling the contextual memory both for preserving or resetting it as situated to the on-going task processes.

How can the network acquire such a sophisticated mechanism involving with a control of the contextual memory? We assumed that such contextual control mechanism could be developed if the network is tutored for longer iterations of the sessions. In order to test this assumption, we conducted learning experiment in which the network was trained with 26 trials each of which consists of 9 continued sessions. The experiment task adopted a simplified form of imperative gesture consisting of 3 concatenated movement primitives (12 different primitive sequences were arbitrarily selected), followed by a forward order command and 3 different speed commands. All of the possible 36 gesture sequences were included in the training.

After the training the network under the aforementioned conditions, the performance of the robot response was examined by proceeding 36 continued sessions in which the gesture sequence was randomly selected from 36 possible gesture sequences at each session. The test results showed that the robot performed with the success rate of 84.4%. This result was compared with the condition of training with the single session condition where 36 different single session tutoring with different gesture sequence for each was conducted for training the network. After training with the single

session condition was completed, the robot performance was tested while the session was iterated for 36 times continued sessions. It turned out that the success rate of the robot response dropped down less than 10% after the 2nd session although the success rate in the first session was more than 90%. These results imply that the robot can perform successfully for continuing sessions when the network has been trained under the same condition of continued sessions.

For the purpose of investigating the mechanism developed via training with continued sessions, we examined the time development of the slow context unit activity observed during the iterations of sessions in the test trials. Figure 8 shows examples of two different test trials for continued sessions with showing the human imperative gesture patterns at the top row, the robot motor response patterns at the middle row and the slow context activity in the bottom row.

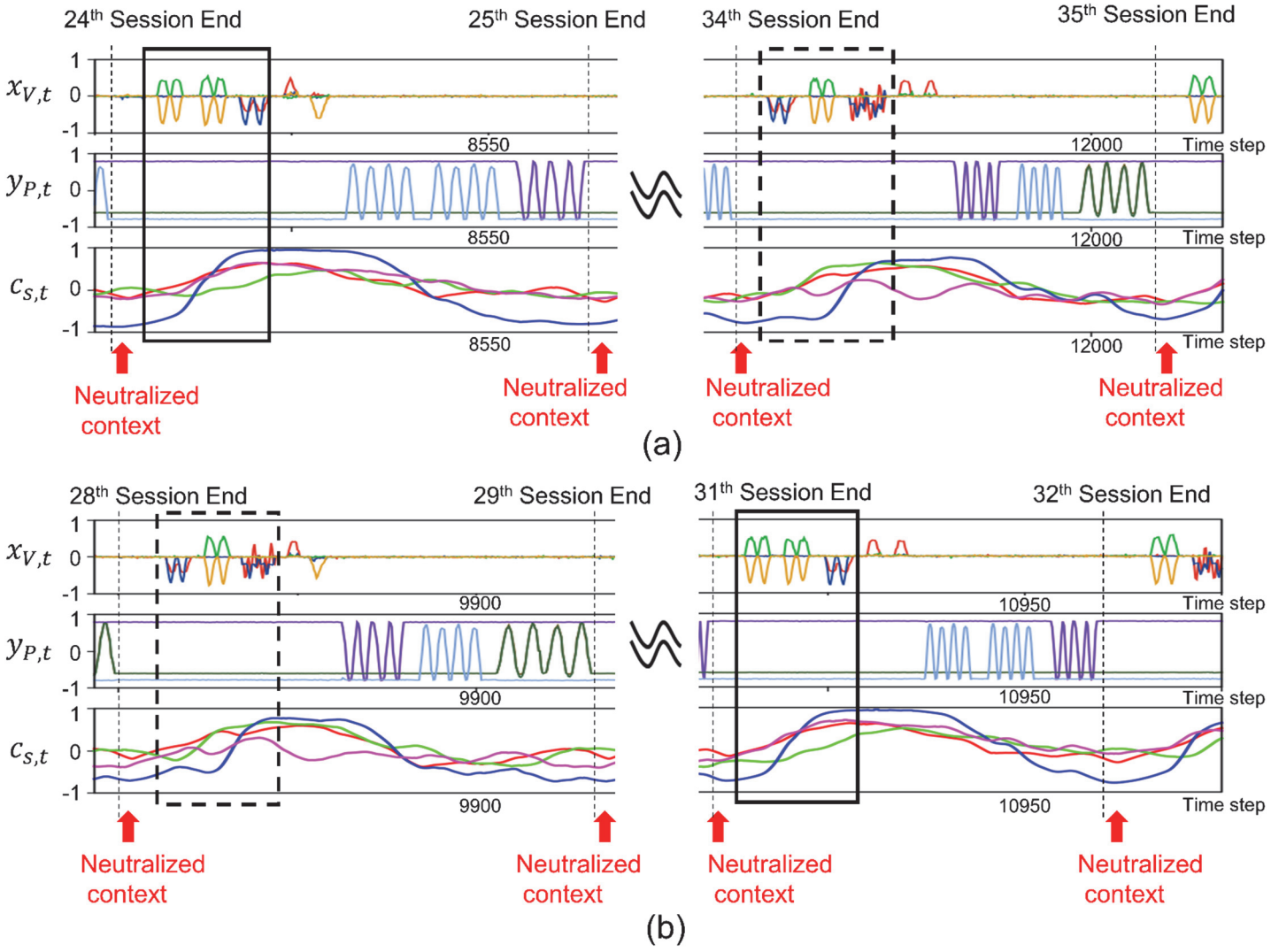

Figure 8. Two examples of test trial results during iterative cycles of sessions plotted with slow context activities. $x_{v,t}$, $y_{p,t}$, and $c_{s,t}$ indicate vision input, proprioception output and fast context. Two solid line rectangles and two dashed line rectangles indicate portions of the same imperative gestures. At each end of session, the context activity becomes similar as pointed by red arrows of "neutralized context".

It can be seen that the slow context activation profile became quite similar when the same imperative gestures were demonstrated regardless of the task content in the previous session (see the context activity profiles in two solid line rectangles for the same category of imperative gesture and in two dashed line rectangles for another same category of imperative gesture). We observed that this finding repeated in the same manner for all other imperative gesture cases. Furthermore, we examined the values of the slow context activity at the end of each session. It was found that the context activity took quite similar values as seen in Figure 8 as indicated by red arrows of "neutralized context". It looks like as if the slow context activity were reset to a neutral value immediately before

the onset of every new session.

In order to confirm the aforementioned observation more quantitatively, variances of the internal state of the slow context units at the end of session among all different sessions were computed for the two conditions of training with a single session and 9 continued sessions. It turned out that the variances were obtained as 0.42 for the condition of training with single session and 0.0645 for the condition of training with 9 continued sessions.

To provide further statistical evidence, in each of the two training conditions, variances in the changes of the internal states of the context units in slow dynamics subnetworks in the Case-A, after encountering all slightly different gesture patterns belonging to the same category of gesture and in the Case-B, after encountering all completely different gesture patterns belonging to different category of gesture are computed. The same analysis method, adopted in Experiment 2 for analyzing the behaviors of the network dynamics, was used to examine the variances in changes of slow context units being compared for the two different cases (the variance in changes of the internal states of the context units in the Case-B divided by the one in the Case-A). For the two training conditions, when training with a single session and 9 continued sessions, the average rates of the variances upon encountering different category of gesture patterns and encountering the same ones were 1.11 and 2.17, respectively. The variance in changes of internal state of slow context units in the Case-A, after encountering all slightly different gesture patterns belonging to the same category of gesture were much smaller than the variance when in the Case-B, after encountering all completely different gesture patterns belonging to different category of gesture in the condition of training with 9 continued sessions. On the contrary, such variances in the two different cases were almost the same in the condition of training with a single session. This result means that the network was not able to categorize the gestures in the condition of training with a single session resulted in poor performance of the network.

The network trained with multiple sessions was able to correctly categorize the gestures, and successfully performed the continuously iterative task sessions by developing a cognitive capability of resetting the task session memory at the end of each session. However, because the network trained with a single session was not able to reset the task session context memory at the end of each task session, the network could not correctly categorize gestures in next continued session and therefore exhibited poor performance. Considering these results, it can be said that resetting the context states

was important to perform the continuously iterative task sessions. In conclusion, these results indicate that the mechanism for autonomous resetting of the context states at the end of each session can be developed provided that the network is trained using relatively long iterative cycles of sessions.

5. Discussion

The current experimental study demonstrated that a humanoid robot controlled by a dynamic neural network model can generate corresponding sequential combinations of motor primitive patterns as responses to observation of various sequential combinations of gesture patterns demonstrated by the human subjects as the consequence of iterative tutoring. It was examined how the adopted MTRNN model can achieve this type of development by learning through continuous flow of the perceptual experience obtained during the tutoring. The experimental results revealed that the MTRNN can achieve generalization by learning in both of the lower perception level for robust recognition against variation in the perceptual features and the cognitive level involving with extraction of compositional semantic rules only by using partial training exemplar. It was also observed that the MTRNN can develop another type of cognitive capability for adaptive control of the contextual memory such as preserving it or resetting it as situated to the on-going task process. Our analysis on the experiment results indicated that such cognitive competency ranging from the lower perceptual level to the cognitive level can be developed depending on the network configuration such as the multiple timescale structure and the topological connectivity assigned to a set of subnetworks in the whole network model.

One possible contribution of the current study to the system level neuroscience might be the finding that a network configuration comprised of a set of locally connected subnetworks assigned with multiple timescale property is essential for development of hierarchically organized cognitive functions. Recently, the sizeable amount of human brain imaging data that has been gathered to date has enabled a global map to be created of both static connectivity and dynamic connectivity between all the different cortical areas (Sporn, 2010). Many studies have speculated that the essential cognitive functions might be largely determined by such connectivity among local brain areas and intrinsic neuronal dynamics in these local areas. Kiebel et al. (Kiebel, Daunizeau, & Friston, 2008), Badre and D’Esposito (Badre & D'Esposito, 2009), and Uddén and Bahlmann (Uddén & Bahlmann,

2012) proposed an idea to explain the rostral-caudal gradient of timescale differences by assuming slower dynamics at the rostral side (PFC) including the Frontal Pole and faster dynamics at the caudal side (M1) in the frontal cortex to account for a possible functional hierarchy in the region. Related to this proposal, Soon and colleagues (Soon, Brass, Heinze, & Haynes, 2008) demonstrated that intention for free action is developed in the Frontal Pole accompanied by relatively slow built-up of readiness potential in human fMRI imaging experiments. This idea of the rostral-caudal gradient of timescale differences can be supported by the current results from the synthetic neurorobotics study showing that the hierarchical cognitive mechanism spanning from the lower perceptual competency to the cognitive competency, which is characterized by systematicity and the capability of controlling contextual memory, can be developed more effectively by utilizing the timescale difference among multiple levels of subnetworks, as well as the local connectivity allowed between adjacent levels of subnetworks.

The observed mechanism of controlling the contextual memory is related to the one reported in the modeling studies for the Wisconsin Card Sorting Test (WCST) conducted by Rougier and O'Reilly (Rougier & O'Reilly, 2002) and Maniadakis et al. (Maniadakis, Trahanias, & Tani, 2012). Rougier and O'Reilly (Rougier & O'Reilly, 2002) proposed an adaptive gating operation scheme which can be trained to act on working memory for storing the currently adopted rules in WCST. In this model the neural activation patterns representing the current rules in the working memory can be preserved by closing the gate until the rules are in conflict with the new rule selected by the experimenter. Maniadakis and colleagues (Maniadakis et al., 2012) showed that a simpler continuous-time recurrent neural network (CTRNN) can reconstruct the same mechanism by applying genetic algorithm to CTRNN. It was shown that the current rule can be preserved in slowly changing internal neuron states which can be reset by events of encountering the conflict.

It is interesting to compare the current scheme of not utilizing the prediction error explicitly for recognizing others' intentions such as latent in imperative gesture and the predictive coding scheme (Ito & Tani, 2004; Murata et al., 2014) of explicitly utilizing it. It can be said that the intention of the other can be recognized unconsciously in the former model and consciously in the latter model if it is assumed that such consciousness originates from prediction error by following the thoughts (Tani, 1998, 2009). It has been known that imitation is performed for both meaningless movement without objects and meaningful motor acts with target objects. Rizzolatti and colleagues (Rizzolatti, Fogassi,

& Gallese, 2001) observed in their monkey experiments that mirror neurons in the parietal lobe and F5 fire in the former and the latter case, respectively. They consider that conscious imitation of meaningful motor acts – *i.e.* motor actions performed to reach a target object, may be accompanied by the firing of mirror neurons in F5. Furthermore, unconscious imitation without meaning can be observed in human neonates, while conscious imitation with meaning can be observed in infants after 2 years old. Meltzoff (Meltzoff, 2005) explained the development of imitation mechanism by proposing a hypothesis on a “like me” mechanism which connects the perceptions of others as “like me” and understanding of others’ minds. In the first stage in newborn, innate sensory-motor mapping can generate the aforementioned imitative behaviors by means of automatic response. In the second stage, infants develop an interdependent relationship between their mental state and performed actions through repetition, which thereby affords learning. Finally in the third stage, infants come to understand that others who act “like me” have mental states “like me”. From these accounts, it can be said that the current model explains an unconscious pathway for communicable congruent responses. The current model could be developed toward conscious goal-directed one if incorporated with an error regression scheme for inferring the intention of others. However, implementation of the error regression scheme to neural network models requires additional artificial setting to models such as fixed length temporal windows as described in the previous section. The future study should investigate how these two different mechanisms for recognition of others’ can be arbitrated in a simple network model.

One major drawback of the current study is that the adopted communicably congruent responses of robots are limited to the experiment setting in the closed laboratory environment. For realizing truly open-ended social human-robot interactions in more natural settings, the current tasks of one-way imperative communication from human to robot should be replaced by mutual communicable congruence tasks between the two sides. Another limitation in the current study is that communicative rules to be learned by robots were predetermined by the experimenter. The communicative rules should emerge or adaptively change through mutual interactions as necessary in real social communication situations. In other words, the communicative rules should be invented or modified depending on the outcome of the on-going interaction for the purpose of gaining utilities in both robot and human sides or achieving shared goals between them. This requires furthermore consideration in the models and the task setting as described in the following.

First of all, it would be more natural if robots and human can develop communicative protocols and rules from simple one to complex one in the course of lone term interaction. For realizing this type of developmental learning, the current learning scheme on the network model should be improved such that it can cope with incremental or dynamic learning rather than the off-line batch learning scheme employed in the current model. Although there have been some trials on developmental or incremental learning with using dynamic neural network models (Nishimoto & Tani, 2003), furthermore intensive investigation of the scheme might be necessary in order to assure both of the stability and flexibility in such dynamic learning.

Another important issue is about the capability on generalization by learning of the network model. The current study showed that the MTRNN model can exhibit generalization capability in some extents, for example, more than 95% motor response accuracy can be attained if 2/3 of all possible combination pairs between the gesture patterns and the motor response patterns are trained, as shown in the experiment 2-1. However, the fact that generalization achieved by training with 2/3 of all possible combinations may not be acceptable in the case of dynamic learning of communicative rules via mutual interactions because the amount of communicative interaction experience which can be utilized for learning at each dynamically changing situation is quite limited.

One possible remedy might be to consider that communication with linguistic complexity may not develop substantially especially in the early phase of communicative interactions. This would make sense because the linguistic complexity in the early stage of verbal communication of infants is quite limited although their communicative interactions with caregivers are known to be quite complex by using other modalities of communication channels such as emotion expression (Nadel, 2002). In this aspect, Asada proposed so-called affective developmental robotics (Asada, 2014) in which he assumed that multiple stages of emotional development observed in human infants should play crucial roles also in development of social communications between robots and human. Future study should focus on building of both psychologically and neurobiologically plausible models which can account for possible interactions between development of emotional communication and that of language-like communication characterized by systematicity. The success of such study should contribute significantly to the realization of truly open-ended human-robot social interaction.

Conclusion

The current study investigated how robots controlled by a particular dynamic neural network model can perform communicably congruent tasks, which are comprised of a set of human gesture patterns and corresponding robot motor responses, via tutoring by the experimenter. The experimental results showed that the MTRNN as an adopted model can achieve generalization by learning in the lower perception level for robust recognition against variation in the perceptual features such as speed and amplitude in the demonstrated gesture patterns by using only partial training data. It was shown that generalization can be achieved also at the cognitive level by extracting compositional semantic rules latent in the continuous perceptual patterns experienced during the tutoring. Furthermore, it was observed that the MTRNN can develop another type of cognition capability for controlling of the contextual memory such as for preserving the current context or resetting it as situated to the on-going task process. It was discussed that the current study should be extended furthermore to include mutual communicative interaction as well as dynamic learning scenario for realizing more natural development of human-robots interactions.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP)(No.014R1A2A2A01005491) and by the Industrial Strategic Technology Development Program (10044009) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

Appendix A. Adaptive learning rate algorithm

In each epoch, the learning rate was updated using by the following algorithm (Namikawa, Nishimoto, & Tani, 2011).

(1) Calculate the delta errors using randomly selected λ percent of the training sequences.

(2) Calculate the rate (r) of the KL-divergence before updating parameters, and the KL-divergence after updating the parameters using whole training sequences.

(3) If $r > r_{th}$, then update α to $\alpha\alpha_{dec}$ and go back to (2).

(4) If $r < 1$, then update α to $\alpha\alpha_{inc}$ and go to the next epoch.

In this study, we set r_{th} to 1.1, α_{dec} to 0.7, and α_{inc} to 1.2 based on the parameter setting used by Namikawa (Namikawa et al., 2011).

In each iteration of the training process, by calculating the gradient of the parameter using a only selected λ percent of the training sequences and determining the learning rate based on the r over entire set of training sequences, *i.e.* constraining the network to minimize prediction error of the entirety of training sequences, not just attempt to minimize error of the only the selected sequences when updating its parameters, we could reduce the over-fitting problem. In other words, we could improve the generalization capability of the network. Adjusting the learning rate by only utilizing the training error, however, has a limitation in improving the generalization capability of the network. Because the network encounters only training sequences, it would extract the rules that only applicable to the minimization of error on training sequences, *i.e.* it would become fatally become over-fitted to the training sequences, as training proceeded, and error on test one would start to increase.

In this context, the training data set were divided into two groups. One group consisted of the sequences used for calculating the delta error and adjusting the learning rate. Another group consisted of the sequences that were concerned with only adjusting the learning rate and used as the validation data set. By also considering minimization error on the validation data set in updating the learning rate, the network would be able to indirectly sense the existence of the external “world.” That is because it would encourage the network to acquire skills and extract rules that are generally applicable to the external “world” rather than acquiring specific skills and extracting particular rules only be applicable to the its “world”. The concept of considering the validation error in adjusting the learning rate is similar to Duffner & Garcia’s (Duffner & Garcia, 2007) method. However, the validation data set in the current study included sequences that consisted of the inexperienced human gestures and robot responses, whereas the validation data in their work were only different from training patterns by noise. Consequently, the adaptive learning rate algorithm in this study, which considers validation errors when adjusting the learning rate, has more important meaning in general rule learning, because this algorithm can widen the network’s awareness of the world.

Appendix B. Criteria of best training epoch

The best epoch was selected after 300,000 epochs, while considering both training and validation errors, instead of stopping the training when the validation error starts to increase, because both training and validation errors fluctuated as training proceeded. The epoch that minimizes the following error function (Eq. B.1) was treated as the best epoch. :

$$E = E_{tr} + \frac{N_{tr}}{N_v} E_v \quad \text{(B.1)}$$

where, E_{tr} and E_v are the training and validation errors, and N_{tr} and N_v are the numbers of training and validation data, respectively.

References

Aldebaran. (2014). who-is-pepper. from http://www.aldebaran.com/en/a-robots/who-is-pepper

Arbib, M. A. (2012). *How the brain got language: The mirror system hypothesis* (Vol. 16): Oxford University Press.

Asada, M. (2014). Towards Artificial Empathy. *International Journal of Social Robotics*, 1-15.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., . . . Yoshida, C. (2009). Cognitive developmental robotics: a survey. *Autonomous Mental Development, IEEE Transactions on, 1*(1), 12-34.

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience, 10*(9), 659-669.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in cognitive sciences, 4*(3), 91-99.

Billard, A. (2002). ft Imitation: A Means to Enhance Learning of a Synthetic Protofanguage in Autonomous Robots. *Imitation in animals and artifacts*, 281.

Billard, A., & Dautenhahn, K. (1998). Grounding communication in autonomous robots: an experimental study. *Robotics and Autonomous Systems, 24*(1), 71-79.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 1): springer New York.

Bowie, J. (2008). Proto-discourse and the emergence of compositionality. *Interaction Studies, 9*(1), 18-33.

Breazeal, C. (2004). Social interactions in HRI: the robot view. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 34*(2), 181-186.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *Robotics and Automation, IEEE Journal of, 2*(1), 14-23.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., . . . Nori, F. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *Autonomous*

Mental Development, IEEE Transactions on, 2(3), 167-195.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., & Striegnitz, K. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics*, 133-160.

Clark, A. (1999). An embodied cognitive science? *Trends in cognitive sciences, 3*(9), 345-351.

Cummins, R. (1996). Systematicity. *The Journal of Philosophy*, 591-614.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 679-704.

Demiris, J., & Hayes, G. (2002). f 3 Imitation as a Dual-Route Process Featuring Predictive and Learning Components; 4 Biologically Plausible Computational Model. *Imitation in animals and artifacts*, 327.

Duffner, S., & Garcia, C. (2007). An online backpropagation algorithm with validation error-based adaptive learning rate *Artificial Neural Networks–ICANN 2007* (pp. 249-258): Springer.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1), 3-71.

Fujita, M., & Kageyama, K. (1997). *An open architecture for robot entertainment.* Paper presented at the Proceedings of the first international conference on Autonomous agents.

Gaussier, P., Moga, S., Quoy, M., & Banquet, J.-P. (1998). From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence, 12*(7-8), 701-727.

Iizuka, H., & Ikegami, T. (2004). Adaptability and diversity in simulated turn-taking behavior. *Artificial Life, 10*(4), 361-378.

Ito, M., & Tani, J. (2004). On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system. *Adaptive Behavior, 12*(2), 93-115.

Kelso, J. S. (1997). *Dynamic patterns: The self-organization of brain and behavior*.

Kendon, A. (2004). *Gesture: Visible action as utterance*: Cambridge University Press.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology, 4*(11), e1000209.

Lewkowicz, D. J., & Lickliter, R. (1995). A dynamic systems approach to the development of cognition and action. *Journal of Cognitive Neuroscience, 7*(4), 512-514.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science, 15*(4), 151-190.

Maniadakis, M., Trahanias, P., & Tani, J. (2012). Self-organizing high-order cognitive functions in artificial agents: Implications for possible prefrontal cortex mechanisms. *Neural Networks, 33*, 76-87.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. *Perspectives on imitation: From neuroscience to social science, 2*, 55-77.

Murata, S., Yamashita, Y., Arie, H., Ogata, T., Tani, J., & Sugano, S. (2014). Generation of Sensory Reflex Behavior versus Intentional Proactive Behavior in Robot Learning of Cooperative

Interactions with Others.

Nadel, J. (2002). Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism. *The imitative mind: Development, evolution, and brain bases*, 42-62.

Nadel, J., Revel, A., Andry, P., & Gaussier, P. (2004). Toward communication: first imitations in infants, low-functioning children with autism and robots. *Interaction Studies, 5*(1), 45-74.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science, 15*(4), 211-229.

Namikawa, J., Nishimoto, R., & Tani, J. (2011). A neurodynamic account of spontaneous behaviour. *PLoS computational biology, 7*(10), e1002221.

Nishimoto, R., & Tani, J. (2003). Learning to generate combinatorial action sequences utilizing the initial sensitivity of deterministic dynamical systems *Computational Methods in Neural Modeling* (pp. 422-429): Springer.

Paine, R. W., & Tani, J. (2005). How hierarchical control self-organizes in artificial adaptive systems. *Adaptive Behavior, 13*(3), 211-225.

Park, G., & Tani, J. (2015). *Development of Compositional and Contextual Communication of Robots by using the Multiple Timescales Dynamic Neural Network.* Paper presented at the Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2015 Joint IEEE International Conferences on.

Port, R. F., & Van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*: MIT press.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661-670.

Rosch, E., Thompson, E., & Varela, F. J. (1992). *The embodied mind: Cognitive science and human experience*: MIT press.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive science, 26*(4), 503-520.

RUMELHART, D., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error propagation. *Parallel Distributed Processing*, 318-362.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in cognitive sciences, 3*(6), 233-242.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience, 11*(5), 543-545.

Sporn, O. (2010). Networks of the Brain: MIT Press, Cambridge, Massachusetts.

Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communications Monographs, 60*(4), 275-299.

Sugita, Y., & Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior, 13*(1), 33-52.

Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems

perspective. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 26*(3), 421-436.

Tani, J. (1998). An interpretation of theself'from the dynamical systems perspective: a constructivist approach. *Journal of Consciousness Studies, 5*(5-6), 5-6.

Tani, J. (2003). Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. *Neural Networks, 16*(1), 11-23.

Tani, J. (2009). Autonomy of Self at criticality: The perspective from synthetic neuro-robotics. *Adaptive Behavior, 17*(5), 421-443.

Thelen, E. (1994). Smith, LB (1994). A dynamic systems approach to the development of cognition and action: Cambridge, MA: MIT Press.

Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental science, 9*(2), 125-147.

Uddén, J., & Bahlmann, J. (2012). A rostro-caudal gradient of structured sequence processing in the left inferior frontal gyrus. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 2023-2032.

Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS computational biology, 4*(11), e1000220.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *MultiMedia, IEEE, 19*(2), 4-10.