Symbol emergence by combining a reinforcement learning schema model with asymmetric synaptic plasticity

T. Taniguchi
Dept. of Mechanical Eng. and Sci., Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, JAPAN
tanichu@groove.mbox.media.kyoto-u.ac.jp

T. Sawaragi
Dept. of Mechanical Eng. and Sci., Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, JAPAN
sawaragi@me.kyoto-u.ac.jp

***Abstract*— A novel integrative learning architecture, RLSM with a STDP network is described. This architecture models symbol emergence in an autonomous agent engaged in reinforcement learning tasks. The architecture consists of two constitutional learning architectures: a reinforcement learning schema model (RLSM) and a spike timing-dependent plasticity (STDP) network. RLSM is an incremental modular reinforcement learning architecture. It makes an autonomous agent acquire behavioral concepts incrementally through continuous interactions with its environment and/or caregivers. STDP is a learning rule of neuronal plasticity that is found in cerebral cortices and the hippocampus. STDP is a temporally asymmetric learning rule that contrasts with the Hebbian learning rule. We found that STDP enables an autonomous robot to associate auditory input with its obtained behavioral concepts and to select reinforcement learning modules more effectively. Auditory signals that are interpreted based on obtained behavioral concepts are revealed to correspond to "signs" in Peirce's semiotic triad. This integrative learning architecture is evaluated in the context of modular learning.**

***Index Terms*— Symbol emergence, constructive semiotics, reinforcement learning, schema model, STDP, modular learning.**

I. Introduction

The symbol grounding problem (SGP) is an important topic in the context of research on both human and artificial intelligence [4]. The SGP is concerned with how to ground the symbolic system of an agent. This problem often sets that a symbolic system was given a priori as a premise. However, a symbolic system does not exist out of an autonomous agent – only inside one. Symbolic systems have come to be seen as being closely connected to our embodiment. Therefore, an agent cannot ground its symbolic system, given by a designer, out of the agent. We believe that one's symbolic system has to emerge through interactions with his/her environment and/or caregivers. In other words, we think that the symbol emergence problem (SEP) should take the place of the SGP. The SEP is a radical bottom-up approach to constructing a symbolic system [10–12].

What is a symbol? Clarifying the definition of the word is necessary to cope with the hard problem, the SEP. Peirce, who started *semiotics*, described a symbol as a triadic relationship of a "sign", "object", and "interpretant"(Fig. 1) [2].

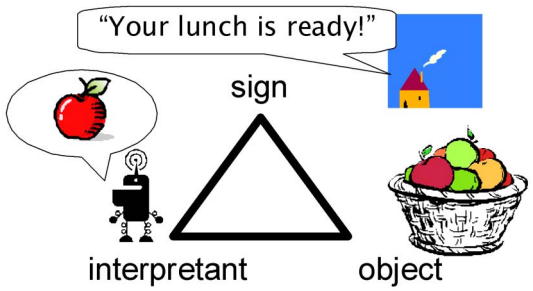

Fig. 1. Peirce's semiotic triad

A simple notion about a symbol is often dyadic. A reference and referent are considered as constituents of a symbol. However, the third participant of a symbol is important in Peirce's semiotics. It is an "interpretant". Usually, no relationship exists between the "sign" and "object" a priori. "Interpretant", which is produced by an autonomous agent, makes them connected. From the viewpoint of semiotics, a symbolic system is not a static system but a dynamic phenomenon.

In Fig. 1, "sign" corresponds to a voice, "your lunch is ready!", and "object" corresponds to a prepared lunch, apples. A robot hears the voice, and an image of an apple is elicited in its mind. It then rushes to a place where the lunch is served. The first interpretant is called an "emotional interpretant", and the second is called an "energetic interpretant". The "energetic interpretant" is a behavior that is evoked by interpreting an incoming "sign". The relationship between the "interpretant" and "sign" is not explicitly given from outside an autonomous agent but is obtained autonomously through interactions with his/her environments and/or caregivers. In this paper, we focus on the "energetic interpretant". To become able to evoke a behavior when referring to a "sign", an autonomous agent has to achiecve two learning processes. One is incremental acquisition of behavioral concepts, and the other is acquisition of the relationship between signs and upcoming situations. If an agent acquires them, he/she can be aware of what kinds of situations will happen in the near future by referring to the sign. After that, the agent can

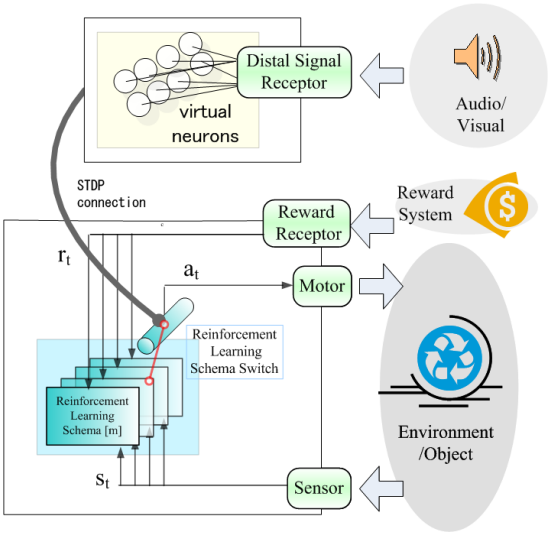

Fig. 2. RLSM with an STDP network

sellect an appropriate behavior immediately. The first learning process is achieved by the reinforcement learning schema model (RLSM) [13], and the second one is achieved by spike timing-dependent plasticity (STDP), which is a synaptic plasticity found in the cerebral cortex and the hippocampus [6]. In this paper, we describe a novel integrative learning architecture that combines these two learning architectures (Fig. 2) .

II. RLSM: Reinforcement Learning Schemata Model

A. Basic concepts

Several modular learning architectures have been proposed. Wolpert et al. proposed MOSAIC as a model of a cerebellum obtaining a multiple internal model [16]. Jacobs et al. proposed a mixture of experts [5]. Singh extended this idea to reinforcement learning and proposed compositional Q-learning (CQL) [8]. However, these modular learning architectures did not achieve on-line incremental modular acquisition.

Taniguchi et al. proposed a dual-schemata model based on Piaget's schema theory [11], [12]. The schema system is defined as an autonomous distributed cognitive system. The schema system explains human infant development especially in the sensorimotor period. They also extended that computational schema model to reinforcement learning and proposed the RLSM as Singh's extended mixture of experts to CQL [13].

The basic concepts of the computational schema model are as follows. The schema model is characterized by two pairs of processes driven by incoming experiences. A schema *assimilates* experiences that are predicted with sufficient accuracy by its inner prediction function. The experiences *accommodate* its inner functions. This cyclic process is called an *equilibration* process. However, if every schema refuses to assimilate a novel experience, a new schema is created for the situation that produces the experience. This process is called *differentiation*. These equilibration and differentiation processes are the basic concepts of the schemata model.

B. Algorithm

RLSM is formulated based on Q-learning [14]. The λ-th schema has two functions: state-action-value function Q^λ and Q^λ's second order statistics function $Q^{(2)\lambda}$. In temporal difference (TD) learning including Q-learning, errors in the value function cannot be observed directly: therefore temporal difference errors must be considered. $Q^{(2)\lambda}$ is considered to estimate Q^λ's standard deviation[1], $\hat{\sigma}^\lambda$. TD-error δ_t and secondary TD-err $\delta_t^{(2)}$ for each λ-th schema are calculated using

$$\delta_t^\lambda = r_t + \gamma V^\lambda(s_{t+1}) - Q^\lambda(s_t, a_t) \ and \tag{1}$$

$$\delta_t^{(2)\lambda} = r_t^2 + 2\gamma r_t V^\lambda(s_{t+1}) + \gamma^2 Q^{(2)\lambda}(s_{t+1}, a^*_{t+1}) - Q^{(2)\lambda}(s_t, a_t), \tag{2}$$

where

$$V^\lambda(s_t) = Q(s_t, a^*_t), \tag{3}$$

$$a^*_t = argmax_a Q^\lambda(s_t, a), \tag{4}$$

and γ is a discount parameter. Each function is updated using these errors:

$$Q^\lambda(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t \tag{5}$$

$$Q^{(2)\lambda}(s_t, a_t) \leftarrow Q^{(2)\lambda}(s_t, a_t) + \alpha\delta_t^{(2)} \tag{6}$$

$$\hat{\sigma}^\lambda = \sqrt{Q^{(2)\lambda} - (Q^\lambda)^2}. \tag{7}$$

By dividing δ_t by estimated standard deviation $\hat{\sigma}_t$, we obtain a dimensionless number, R_t^λ, as subjective error:

$$R^\lambda(t) \equiv |\delta_t^\lambda / \hat{\sigma}_t^\lambda|^2 \tag{8}$$

$$\breve{R}^\lambda(t + \triangle t) = (1-p)R^\lambda(t) + p\breve{R}^\lambda(t) \tag{9}$$

$$\sim \int_0^\infty \frac{1}{\tau} exp(-\frac{s}{\tau}) R^\lambda(t-s) ds \tag{10}$$

$$= \int_{-\infty}^\infty \mathbf{W}_- R^\lambda(t+s) ds \tag{11}$$

$$\mathbf{V}^\lambda(t) = \chi_c(n_p \breve{R}^\lambda(t), n_p) \tag{12}$$

$$n_p = \frac{1+p}{1-p} \tag{13}$$

$$\mathbf{W}_-(t) = \begin{cases} 0 & if\ t > 0 \\ \frac{1}{\tau}\exp(-|t|/\tau) & otherwise \end{cases},$$

where $\chi_c(x, n)$ is a chi-squared one-sided cumulative function, $P(X > x)$, and $\breve{R}^\lambda$ is a temporal weighted average of the subjective errors. $\breve{*}$ is an operater defined as eq. 11. $\mathbf{V}^\lambda$ is the λ-th schema activity. "Schema activity" means

[1]The second order statistics of a value function are also considered in Bayesian Q-learning [3].

how near the robot's facing environment and reward function are to the λ-th reinforcement learning schema. The p is a parameter representing how long a schema retains a previous recognition. In continuous time, time constant $\tau = \triangle t/(1-p)$ corresponds to p; $\triangle t$ is the continuous time for one step in discrete time. Furthermore, the calculation of $\breve{R}^{\lambda}$ can be rewritten by using window function $\mathbf{W}_{-}$ as eq. 11. A reinforcement learning schema decides whether to assimilate an experience or to reject it by referring to schema activity $\mathbf{V}^{\lambda}$. If all the schemata reject an incoming experience, differentiation is initiated, and a new schema is created. This algorithm enables an autonomous robot to notice qualitative changes in a time series of s_t, a_t, and r_t and to obtain novel behavioral concepts incrementally. Significance parameter $\mathbf{V}_{\alpha}$ is set, and the probability, $P(\lambda)$, with which λ-th schema is selected, is defined as

$$\mu(H^{\lambda}) = sgn(\mathbf{V}^{\lambda}(t) - \mathbf{V}_{\alpha}) \quad (14)$$

$$P(\lambda) = \mu(H^{\lambda}) \prod_{k=0}^{\lambda-1} (1 - \mu(H^{k})), \quad (15)$$

where

$$sgn(x) = \begin{cases} 1 & if \quad x > 0, \\ 0 & otherwise \end{cases} \quad (16)$$

and $\mu(H^{\lambda})$ is the truth value of hypothesis H^{λ}, which means that the robot's facing environment corresponds to the λ-th schema. H^{0} is a dummy hypothesis defined to simplify the equation ($\mu(H^{0}) = 0$).

III. STDP: Spike Timing-Dependent Plasticity

The STDP learning rule was found to operate in the cerebral cortex and hippocampus [6]. Before the finding, the Hebbian learning rule was considered to play the main role in associative learning, which these brain areas are considered to be engaged in. What is a crucial difference between the STDP rule and the Hebbian rule? The STDP rule is a temporally asymmetric rule, but the Hebbian rule is temporally symmetric (Fig. 3). However, the functions that STDP has have not been revealed. In this section, we describe a modified STDP rule that makes time delays in modular selection much shorter.

A. STDP learning rule

Various studies on computational STDP learning rules were conducted after neurobiological experiments found the asymmetric learning rule. The most common rule [7] is described as

$$\triangle w = \sum_{i,o} W(w, t_o^{post} - t_i^{pre}) \quad (17)$$

$$W(w, \triangle t) = \begin{cases} \frac{S_+(w)}{\tau_+} \exp(-|\triangle t|/\tau_+) & if \ t > 0 \\ -\frac{S_-(w)}{\tau_-} \exp(-|\triangle t|/\tau_-) & otherwise \end{cases}$$

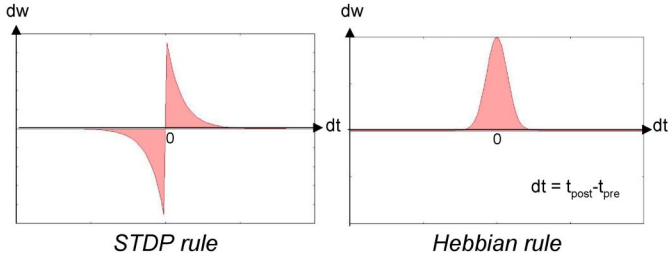

Fig. 3. Difference between asymmetric STDP learning rule and symmetric Hebbian learning rule

where $\triangle w$ is the change in the synaptic weight between the pre and post neuron, t_o^{post} is the o-th time when the post neuron fires, and t_i^{pre} is the i-th time when the pre neuron fires. $W(w, \triangle t)$ is an asymmetric function with respect to $\triangle t$ as shown in Fig. 3. τ_+ and τ_- are the time constants, and S_+ and S_- are the learning gain in the STDP. Generally, these gain parameters S_+ and S_- depend on synaptic weight w. However, we treat only the case where the gain parameters are independent of the synaptic weight in this paper. In addition to that, we set time constants $\tau_+ = \tau_- = \tau$ symmetrically.

If orderly input-output spikes were given to a STDP synaptic connection without other additional learning rules, the synaptic weight would diverge. Therefore, a hard boundary [9], which restrict w between the minimum and maximum value, or a soft boundary [1], which adds an additional multiplicative rule to the additive STDP rule, should be designed. We introduce a multiplicative decay term in proportion to w.

$$\triangle w = \sum_{i} -S_{decay} w(t_i^{pre}). \quad (18)$$

S_{decay} is a gain parameter of the decay term. By using the operater $\breve{*}$, The total modified STDP rule can be transformed into a simple dynamical system

$$\dot{w} = S_+ I \breve{O} - S_- O \breve{I} - S_{decay} w I, \quad (19)$$

where I and O are input and output function, respectively. Each function is defined as

$$I(t) = \sum_{t_i \in \mathcal{T}_I(T)} \delta(t - t_i) \quad (20)$$

$$O(t) = \sum_{t_o \in \mathcal{T}_O(T)} \delta(t - t_o) \quad (21)$$

where $\mathcal{T}_I(T)$ and $\mathcal{T}_O(T)$ are the groups of time when the pre neuron fires and the post neuron fires, respectively, until time T. Dirac's $\delta(t)$ represents a spike.

B. What does STDP encode to synaptic weight?

RLSM selects the reinforcement learning schema, i.e., the Q-table, based on $\breve{R}^{\lambda}(t)$ defined in eq. 11. However, the

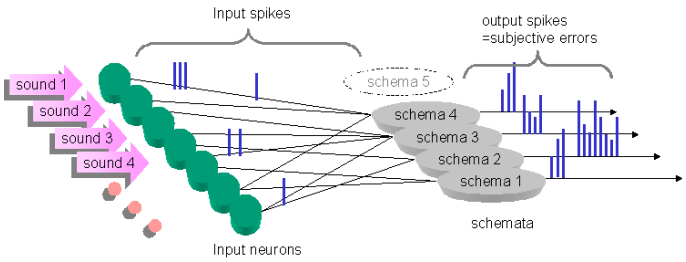

Fig. 4. STDP neuronal network connected to reinforcement learning schemata firing spikes that encode subjective errors

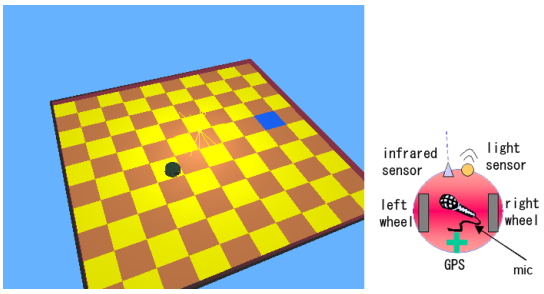

Fig. 5. Left: simulation space; right: Khepera's sensory-motor system

window function $\mathbf{W}_-$ has its center of gravity at $t = -\tau$. This means that the schemata are selected based on past subjective errors. MOSAIC also has this kind of time delay. To reduce this time delay, τ should be smaller. However, the modular selection and organization would become unstable if τ were too small. Therefore, τ has a trade-off regarding the time delay of modular selection and stability in modular organization.

In contrast with the past window function $\mathbf{W}_-$, We define a future window function $\mathbf{W}_+$ as

$$\mathbf{W}_+(t) = \begin{cases} \frac{1}{\tau}\exp(-|t|/\tau) & if\ t > 0 \\ 0 & otherwise \end{cases}$$

which has a center of gravity at $t = \tau$. Next, we try to estimate the future averaged subjective error $\bar{R}$ under the condition a pre neuron fires.

$$E[\bar{R}|t_i \in \mathcal{T}_I] \quad (22)$$

$$= E[\int_{-\infty}^{T} \mathbf{W}_+(s)R(t_i+s)ds|t_i \in \mathcal{T}_I] \quad (23)$$

$$\sim E[\int_{-\infty}^{T} (\mathbf{W}_+(s) - \mathbf{W}_-(s))R(t_i+s)ds|t_i \in \mathcal{T}_I] + \breve{R}(t)$$

By transforming the first term, we obtain

$$\sim \frac{\iint_{-\infty}^{T} I(t)(\mathbf{W}_+(s-t) - \mathbf{W}_-(s-t))R(s)dsdt}{\#(\mathcal{T}_I(T))}$$

$$= \frac{\int_{-\infty}^{T} R\breve{I} - I\breve{R}dt}{\int_{-\infty}^{T} Idt}. \quad (24)$$

By differentiating this formula with respect to T, we obtain

$$\dot{w} = \frac{1}{\#(\mathcal{T}_I(T))}(R\breve{I} - \breve{R}I - wI) \quad (25)$$

$$= S_+ I\breve{R} - S_- R\breve{I} - S_{decay}wI, \quad (26)$$

where $S \equiv S_+ = S_- = S_{decay} = 1/\#(\mathcal{T}_I(T))$. This shows the modified STDP learning rule can encode the difference between the estimated future averaged subjective errors and past averaged subjective errors to the synaptic weight by considering R as output spikes of the post neuron.

C. Integrative learning architecture for Symbol Emergence

We replace the post neurons in the STDP network by the reinforcement schemata firing R^λ as output spikes (Fig. 4). The reason is as we mentioned in the previous subsection; the STDP network automatically encodes the information of the estimated $\bar{R}^\lambda$ to the synaptic weight w_i^λ between the i-th pre neuron and the λ-th schema. For this paper, we assumed that the pre neuron fires when the sound characteristics of each neuron come in. This integrative learning architecture has both the STDP learning rule and RLSM learning rule running. Under these conditions, the architecture can make sense out of incoming sounds that are related to changes in schemata activities. The window function $\mathbf{W}_+$ has its center of gravity at $t = \tau$. Therefore, the RLSM becomes able to select an appropriate schema by referring to the incoming sound interpreted as a "sign" of the change that will happen in the future. The λ-th schema's inner state Ψ^λ changes as

$$\dot{\Psi}^\lambda = \sum_i (w_i^\lambda - \Psi^\lambda) sgn(w_i^\lambda - w_{dz}^\lambda) I_i(t) - \frac{1}{\tau}\Psi^\lambda, \quad (27)$$

where w_{dz}^λ denotes the dead zone of the synaptic weight. w_{dz}^λ is defined as $w_{dz}^\lambda = k\sqrt{2var[R^\lambda]S_i^\lambda}$. k is a constant, and S_i^λ is a gain parameter in the STDP learning rule. The first term reflects the synaptic weight w_i^λ of the nearest input spike to the λ-th schema's inner state. The second term denotes the decay term of the inner state. By using this inner state, the RLSM can select an appropriate schema more effectively without any wasteful time delays.

IV. Evaluation

We evaluated the RLSM by using a 2D simulation for the Khepera mobile robot [15].

A. Conditions

The simulation space and a Khepera's sensory-motor system are illustrated in Fig. 5. We used Webots, produced by cyberbotics, to simulate Khepera's dynamics. The square simulation space was enclosed by walls 2 m long and 10 cm high. A light source was located at a height of 10 cm at the center of the space. Khepera has two wheels as

Fig. 6. Reward functions

a motor system, and their rotational velocities can be set independently at each time step. Khepera's sensory system is comprised of an infrared sensor, a light sensor, and a GPS. Q-learning usually requires a discrete state space. Therefore, we divided Khepera's x,y coordinates and its angle of direction, obtained from the GPS, into six parts, and we defined $216 (= 6 \times 6 \times 6)$ states. The action space was also made discrete by defining five representative motor outputs: i.e., forward, back, right, left, and stop. Forward and back move the robot about 30 cm per step, and right and left rotate it about 60° per step. The values of the infrared sensor (ds) and the light sensor (ls) were limited to between 0 and 1. They were used only to calculate the rewards. To catch incoming sounds, a mic sensor was also provided, and it was connected to input neurons, which are modeled in Fig. 4.

We prepared three reward functions: $r^1 = ds$, $r^2 = 1.5ls$, $r^3 = 1.8v_{forward}$, where $0 \leq v_{forward} \leq 1$ is the value given when Khepera advances. These reward functions mean that the caregiver wants the robot to face a wall, to gaze at a light, and to run around, respectively (Fig. 6). The parameters for reinforcement learning were set to $\alpha = 0.2$ and $\gamma = 0.8$. The parameters for the schemata model were set to $p = 0.999$ and $\mathbf{V}_\alpha = 0.0001$. The parameter for the STDP was set to $k = 0.75$. In our previous work [13], we investigated whether the RLSM enabled Khepera to obtain several behavioral concepts, i.e., reinforcement learning schemata, and to recall them while it was interacting with the environment. The reward functions in this environment were switched in turn. Each trial consisted of 200 steps. The reward functions were set to r^1 for $(0 < trial \leq 750)$, r^2 for $(750 < trial \leq 1500)$, and r^3 for $(1500 < trial \leq 2250)$. Subsequently, each reward function was used in turn for 250 trials.

As shown in Fig. 7, each schema activity, $\mathbf{V}^\lambda$, had a successful transition. Also, the schema, initially only one, differentiated into three. Finally, three behavioral concepts corresponding to the three reward functions were organized. Each schema was selected as shown in Fig. 7.

However, in this experiment, a 2000 step had already passed on average when RLSM recalled an appropriate schema in each switch of reward functions. To overcome this time delay problem, Khepera should become able to use sounds that ring when the corresponding reward functions are adaptively selected as “signs”. By using the STDP learning rule, Khepera can learn the relationship between sounds and

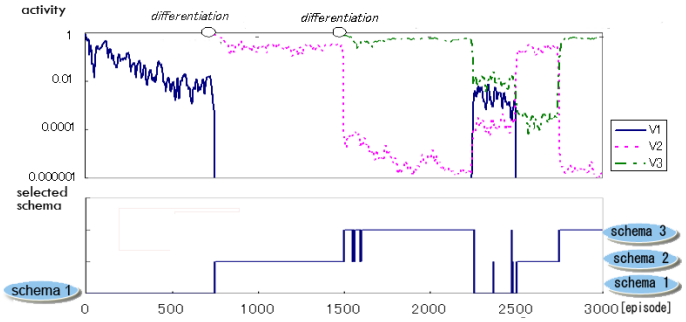

Fig. 7. Top: schema differentiation process and transition of schema activities; bottom: selected reinforcement learning schema

the next schema that should be activated, and it should recall an appropriate schema effectively. This semiotic process corresponds to the one shown in Fig. 1. We prepared 36 kinds of sounds in the simulation space to evaluate the STDP learning rule. After the three schemata were acquired, we continued to switch the reward functions in turn for each 25 trial. The sounds 1, 2, and 3 were rung when r^1, r^2, and r^3 were selected, respectively. In addition to that, the sounds named from 4 to 36 were rung randomly as noise. Under this noisy condition, an autonomous robot had to determine meaningful sounds that related to its organized schemata. If the robot could determine the relationships, the robot was able to exploit the sounds as “signs”. In the situation, “sign” “object”, and “energetic interpretant” correspond to “sound”, “transition to a situation where a particular reward function is ready”, and “behavior obtained in a schema”, respectively.

B. Results

As shown in Fig. 8, synaptic weights encoded the relationship between the sounds and each schema. In the figure, the synaptic weights around third and fourth sound which have no relationship to schema 1, converged into 0. However, the synaptic weight around the second sound obtained a large value because schema 1 started to output big subjective errors when its facing reward function was switched from r^1 to r^2. The synaptic weight around the first sound obtained a large negative value because schema 1 reduced its subjective errors when a reward function was switched from r^3 to r^1. This is in contrast to the second sound.

By exploiting these obtained synaptic weights, Khepera became able to interpret incoming “signs” to decide which schema to select next. As shown in Fig. 9, time delays that were required when Khepera switched its schema corresponding to external rewards decreased gradually as it learned and exploitied incoming “signs”. This dynamic processes of interpretation might satisfy the conditions of Peirce's semiotic triad. Therefore, we conclude that “symbols” emerged in this experiment.

From the viewpoint of behaviorism, this experiment looks like experiments of operant conditioning. However, this un-

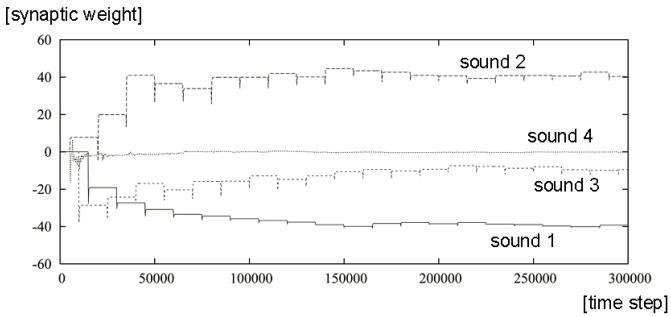

Fig. 8. Transition of synaptic weight between input neurons and schema 1

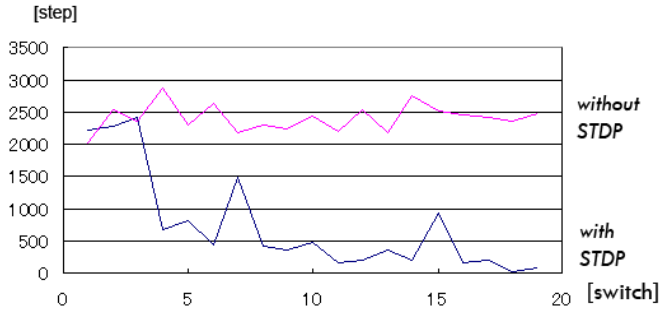

Fig. 9. Course of time delays in switching schemata

derstanding of the experiment is based on the observer's viewpoint. If the observer's viewpoint is set inside the autonomous agent, this learning process will look a positive symbol organization process rather than a passive stimuli-response relationship.

V. Summary

We described a novel integrative learning architecture that consists of a RLSM and an STDP neuronal network. These network achieve symbol emergence in a self-enclosed autonomous agent. Using the RLSM enables an autonomous agent to obtain several learning modules, called reinforcement learning schemata, without any explicit indication except for sensor vectors, motor vectors, and rewards. After obtaining the schemata, using STDP enables the agent to obtain the relationships between incoming sounds and the agent's obtained schemata and to exploit them as "signs" representing future situations. It is important that neither a RLSM learning rule nor a STDP learning rule is supervised-learning rules. Therefore, the integrative learning rule is classified into self-organizational learning rules. The incoming sounds have no meaning before the agent determines their meanings. The emerging triadic relationship of "sign", "object", and "interpretant" was nothing but a "symbol" (Fig. 1). We conclude that our integrative learning architecture achieves symbol emergence.

Acknowledgments

This work was supported in part by the Center of Excellence for Research and Education on Complex Functional Mechanical Systems. (The 21st Century COE program of the Ministry of Education, Culture, Sports, Science and Technology, Japan).

References

[1] L.F. Abbott and S.B. Nelson. Synaptic plasticity: taming the beast. *nature neuroscience supplement*, 3:1178–1182, 2000.
[2] D. Chandler. *Semiotics the BASICS*. Routledge, 2002.
[3] R. Dearden et al. Bayesian q-learning. In *AAAI-98*, pages 761–768, 1998.
[4] S. Harnad. The symbol grounding problem. *Physica D*, 42:35–346, 1990.
[5] R. A. Jacobs, M. I. Jordan, et al. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
[6] H. Markram et al. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *SCIENCE*, 275:213–215, 1997.
[7] Y. Sakai, K. Nakano, and S. Yoshizawa. Synaptic regulation on various stdp rules. *Neurocomputing*, 58–60:351–357, 2004.
[8] S.P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning archive*, 8(3-4):323–339, 1992.
[9] S. Song et al. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *nature neuroscience*, 3:919–926, 2000.
[10] S. Takamuku, Y. Takahashi, and M. Asada. Lexicon acquisition based on behavior learning. In *Proceedings of the 2005 4th IEEE International Conference on Development and Learning*, 2005.
[11] T. Taniguchi and T. Sawaragi. Design and performance of symbols self-organized within an autonomous agent interacting with varied environments. In *IEEE International Workshop on RO-MAN proceedings in CD-ROM*, 2004.
[12] T. Taniguchi and T. Sawaragi. Self-organization of inner symbols for chase: Symbol organization and embodiment. In *IEEE International Conference on SMC 2004 proceedings in CD-ROM*, 2004.
[13] T. Taniguchi and T. Sawaragi. Incremental acquisition of behavioral concepts through social interactions with a caregiver. In *Artificial Life and Robotics (AROB 11th '06) proceedings*, 2006.
[14] C. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8:279–292, 1992.
[15] Webots. http://www.cyberbotics.com. Commercial Mobile Robot Simulation Software.
[16] D.M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11:1317–1329, 1998.