Multimodal joint attention through cross facilitative learning based on μX principle

Yuichiro Yoshikawa*, Tsukasa Nakano†, Minoru Asada*†, and Hiroshi Ishiguro*†
*Asada Synergistic Intelligence Project, Erato, JST
2-1 Yamadaoka, Suita, Osaka, Japan 565-0871
Email: {yoshikawa,ishiguro,asada}@jeap.org
†Dept. of Adaptive Machine Systems, Graduate School of Enginering, Osaka University
Email: {tsukasa.nakano, ishiguro, asada}@ams.eng.osaka-u.ac.jp

***Abstract*—Simultaneous learning of multiple functions is one of the fundamental issues not only to design intelligent robots but also to understand human's cognitive developmental process since we, human, do so in our daily lives but we do not know how to do. Drawing an analogy to the well-known bias in child language development, we propose the *mutual exclusivity selection principle* (μX principle) for learning multi-modal mappings: selecting more mutually exclusive output leads experiences to make underdeveloped complementary mappings more disambiguated. The μX principle is applied to multi-modal joint attention with utterances for lexicon acquisition, and synthetically modeled in both intra- and inter-module levels of output. Through the series of computer simulations, the effects of the μX principle on the mutual facilitation in learning multi-functions and robustness against errors in segmentation of observation are analyzed. Finally, the correspondence of the synthesized development to infant's one is argued based on the simulation with careful behavior by a caregiver.**

I. Introduction

Joint attention is one of the bases for communication, and therefore has been a central topic in developmental psychology [1], [2]. It has been reported that infants gradually come to follow the gaze direction of adults toward wider regions from about 12 to 18 months of age [1]. They seem to perform not only through a single modal observation of the other's attention, i.e., looking at her gaze direction, but also through another modality such as listening to her utterances. However, the coordination among these different modalities does not seem to have matured in the early period of development: infants cannot refer to the gaze direction of adults when they learn word labels of objects from the adults' utterances until about 18 months of age [2] although they have already started acquiring word labels at this age [3]. What kinds of mechanisms successfully constrain such a complex interaction of the developmental processes of multimodal functions for joint attention has been a formidable question.

On the other hand, learning multi-functions is one of the fundamental issues for intelligent robots. How does learning one function affect another one: is there a mutual effect of learning acceleration, is it neutral, or is there deceleration? It seems a feasible approach to start from reproducing the developmental process of an infant to acquiring multimodal joint attention as an example task. Such a study is expected to contribute not only to establishing design principles of intelligent robots but also to synthetically modeling the developmental process of joint attention in human infants as advocated in cognitive developmental robotics [4]. It has shown that a robot can acquire gaze-following only through statistically mapping the caregiver's gaze and locations when it finds something salient [5], [6]. The statistical mapping approach has also been adopted for modeling the development of word-to-object mapping [7], [8], which could be utilized for word-driven joint attention. However, these studies have focused only on either modality. On the other hand, gaze-driven joint attention has been shown to be necessary for statistical learning of word-to-object mapping [9] as infants older than 18 months of age do. However, the simultaneous learning of multimodal joint attention has not been addressed. In considering multi-functional development where functions are complementary to each other, such as gaze-driven and word-driven joint attention, the learning progress of one function can facilitate that of the other.

It has been reported that children exhibit mutually exclusivity bias in language acquisition, i.e., a tendency to associate novel word with novel object [10]. Generalizing it for a developmental mechanism of multimodal mappings, *mutual exclusivity selection principle* (hereafter the μX principle) is introduced: selecting more mutually exclusive output leads experiences to make underdeveloped complementary mappings more disambiguated. The μX principle is considered in both intra- and inter-module levels of output although a previous work on visuo-tactile binding has considered it in the level of mapping only [11]. In the intra-module level, mapping representing the correlation between discrete representations is used to calculate output not only reflecting the correlation but also highlighting the more mutually exclusive correlation. On the other hand, in the inter-module level, the outputs from multi-functions are integrated through weighting each output according to its mutual exclusivity. We expect such biases successfully constrain the simultaneous learning process of complementary functions because mutually exclusive output of a module can be expected to disambiguate the output of the other complementary module.

In this paper, we formalize the learning process of multimodal joint attention as the problem for mutually facilitative learning of complementary functions based gaze-to-location

mapping and word-to-object mapping. After describing how we implement the μX principle into the multimodal attentional system, we report results from computer simulation in three experimental settings. In the first experiment, the effect of μX principle to mutually facilitate learning of multi attentional module are analyzed in detail assuming that the learner's segmentation of the observation is perfect. Then, the segmentation process is assumed to be imperfect in the second experiment to examine the robustness against such errors. In the third experiment, more plausible behavior of the caregiver that promotes the learning process is considered to highlight that the mutually facilitative effects occur in a cross-modal manner where matured parts of each mapping promote learning of immature parts of the other mapping. We finally argue the correspondence of the simulation to the infant development.

II. Multimodal joint attention based on the μX principle

A. Assumptions

Here, we focus on two types of joint attention (hereafter JA), namely gaze-driven JA and word-driven JA, to formalize the problem of multimodal JA. Suppose that there are several objects around a learner and a caregiver, and that the learner tries to acquire JA with the caregiver (see Fig. 1). In every trial of JA, the caregiver tells a word label of an object that the caregiver is looking at. The learner tries to look at the same object without any *a priori* knowledge about the relationships between the caregiver's face pattern and the directions of her gaze and between the caregiver's word labels and the object views named by the labels. To perform multimodal JA, these relationships should be found and utilized for locating an object that the caregiver is looking at.

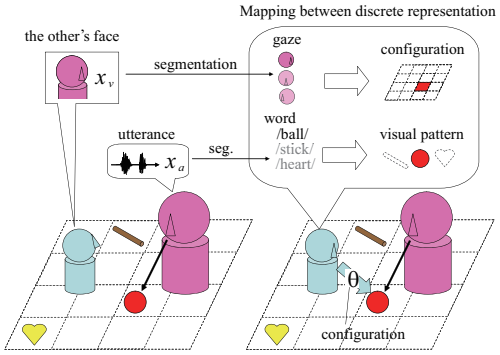


Fig. 1. Situation of learning multimodal joint attention

In order to acquire JA, the learner obtains its configuration $\boldsymbol{\theta} \in \Re^{N_\theta}$ that matches with a location (ex., joint angles or positions in the environment) from visual observation $\boldsymbol{x}_v \in \Re^{N_v}$ and auditory one $\boldsymbol{x}_a \in \Re^{N_a}$, where N_θ, N_v and N_a are the dimension of configuration, visual and auditory observations, respectively. We assume that the learner has already acquired some kinds of discrete representations of the world. Segment vectors, $\bar{\boldsymbol{r}} \in \Re^{M_r}$, $\bar{\boldsymbol{g}} \in \Re^{M_g}$, $\bar{\boldsymbol{w}} \in \Re^{M_w}$, and $\bar{\boldsymbol{p}} \in \Re^{M_p}$ denote the discrete representations of configuration, the other's gaze direction, the word label, and visual pattern, respectively. M_r, M_g, M_w, and M_p denote the numbers of discrete representation for these referents. The i-th element of each segment vector indicates to what extent the referent can be represented by the i-th discrete representation. The segmentation functions ${}^{\bar{g}}\boldsymbol{S}_v : \Re^{N_v} \rightarrow \Re^{M_g}$ and ${}^{\bar{w}}\boldsymbol{S}_a : \Re^{N_a} \rightarrow \Re^{M_w}$ that extract $\bar{\boldsymbol{g}}$ and $\bar{\boldsymbol{w}}$ from the observation are assumed to be given and might not be perfect (see Fig. 2 for an example of the segment vector of the other's gaze direction). Note that, hereafter, a bold character denotes continuous multidimensional values while a bold character with bar on it denotes a set of likelihoods of discrete representation for a certain object.

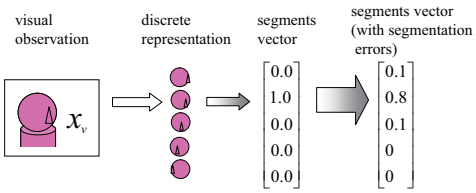


Fig. 2. Example segment vectors of the discrete representation for gaze in the cases without and with errors in segmentation

B. Multimodal attentional modules

The proposed learning architecture consists of two attentional modules, namely gaze-driven attention based on $\bar{\boldsymbol{g}}$ and word-driven attention based on $\bar{\boldsymbol{w}}$, and an arbitrator to integrate them into $\boldsymbol{\theta}$ (see Fig. 3). Each of these modules learns to output a configuration segment vector, $\bar{\boldsymbol{r}}_g \in \Re^{M_r}$ and $\bar{\boldsymbol{r}}_w \in \Re^{M_r}$ to acquire JA based on the information of gaze and words, respectively. The arbitrator finally outputs $\boldsymbol{\theta}$ by integrating $\bar{\boldsymbol{r}}_g$ and $\bar{\boldsymbol{r}}_w$.

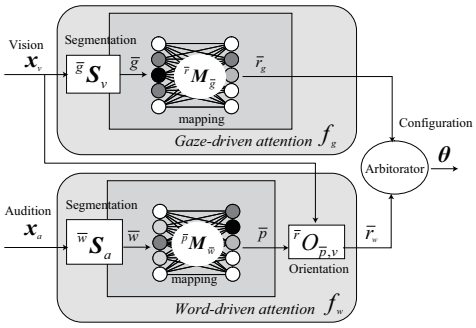


Fig. 3. System of multimodal joint attention

1) Gaze-driven attention: The gaze-driven attentional module is composed by a fixed segmentation function ${}^{\bar{g}}\boldsymbol{S}_v$ and a mapping function ${}^{\bar{r}}\boldsymbol{M}_{\bar{g}} : \Re^{M_g} \rightarrow \Re^{M_r}$ (see the upper box in Fig.3). ${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ maps the discrete representation of observed gaze $\bar{\boldsymbol{g}}$ onto the discrete representation of likely configuration segment vector $\bar{\boldsymbol{r}}_g$ and is updated through the experiences.

${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ is implemented by a two-layered fully connected feedforward network. Each of M_g units in the input layer

encodes the value of each element of $\bar{\boldsymbol{g}}$, and each of M_r units in the output layer encodes one of $\bar{\boldsymbol{r}}$. In other words, the i-th input unit encodes the i-th discrete representation of the other's gaze, and the j-th output unit encodes the j-th discrete representation of configuration. The strength of connection between the i-th input unit and the j-th output one is denoted as ${}^{\bar{r}}\omega_{\bar{g},ij}$ and represents the correlation between them.

Given an input $\bar{\boldsymbol{g}}$, the activation of the j-th output unit ${}^{r}a_{g,j}$ is first calculated as follows: ${}^{r}a_{g,j} = \sum_i^{M_g} {}^{r}\tilde{\omega}_{g,ij}\bar{g}_i$ where a normalized connection strength ${}^{r}\tilde{\omega}_{g,ij}$ is obtained by normalizing ${}^{r}\omega_{g,ij}$. To implement the μX principle, the normalization is given as follows:

$$ {}^{r}\tilde{\omega}_{g,ij} \quad = \quad {}^{r}\omega_{g,ij} \exp\left(-\frac{\sum_{k,k\neq i} {}^{r}\omega_{g,kj}}{\sigma_g^2}\right) \qquad (1) $$

so that the more the j-th output unit is correlated with the i-th input unit and the less it is correlated with other input units, the more the j-th output unit is activated. Note that σ_g^2 is an insensitivity parameter for the mutual exclusivity. Finally, the values of activation are normalized to output $\bar{\boldsymbol{r}}_g$ such as $\bar{\boldsymbol{r}}_g = {}^{r}a_{g,j}/\sum_k {}^{r}a_{g,k}$. Note that the j-th element of $\bar{\boldsymbol{r}}_g$ indicate the likelihood of the j-th representation of the configuration to be selected for performing JA when considering only the information of the caregiver's gaze direction.

2) Word-driven attention: The process of the word-driven attentional module is divided into visual pattern retrieval and orientation toward the retrieved visual pattern. The function of visual pattern retrieval is composed by a fixed segmentation function ${}^{\bar{w}}\boldsymbol{S}_a$ and a mapping function ${}^{\bar{p}}\boldsymbol{M}_{\bar{w}} : \Re^{M_w} \rightarrow \Re^{M_p}$ (see the bottom box in Fig.3). ${}^{\bar{p}}\boldsymbol{M}_{\bar{w}}$ maps the discrete representation of observed utterance $\bar{\boldsymbol{w}}$ onto the discrete representation of likely visual pattern segment vector $\bar{\boldsymbol{p}}$ and is updated through the experiences.

A similar network used for ${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ is also adopted for ${}^{\bar{p}}\boldsymbol{M}_{\bar{w}}$ to map $\bar{\boldsymbol{w}}$ onto $\bar{\boldsymbol{p}}$. Each of M_w units in the input layer encodes the value of each element of $\bar{\boldsymbol{w}}$, and each of M_p units in the output layer encodes one of $\bar{\boldsymbol{p}}$. $\bar{\boldsymbol{p}}$ is calculated from the input $\bar{\boldsymbol{w}}$ through a process similar to that described in Eq. (1) with an insensitivity parameter for mutual exclusivity σ_w^2.

Finally, the visual orientation mechanism ${}^{r}\boldsymbol{O}_{\bar{p},v}$ outputs $\bar{\boldsymbol{r}}_w$ to fixate a region that includes an image similar to the retrieved image. We suppose that ${}^{\bar{r}}\boldsymbol{O}_{\bar{p},v}$ are also given as a normal visual tracking method. Practically, the implementation is described by using the representation of segment vector as follows:

$$ \bar{r}_{w,j} \quad = \quad \begin{cases} s_j & \text{if } s_j > 0 \\ (1 - \sum_j^{M_\theta} s_j)/N_s & otherwise \end{cases} \qquad (2) $$

where $s_j = \sum_k^{M_\theta} \bar{p}_k e_{jk}$ indicates the total word-induced saliency in the visual observation $\boldsymbol{x}_v$ on the j-th configuration while e_{jk} indicates the existence of the k-th visual representation of object there; output is 1 if it exists, otherwise 0. $\bar{r}_{w,j}$ is the j-th element of $\bar{\boldsymbol{r}}_w$ and indicates probabilities to be selected for performing JA when considering only the information of the caregiver's utterances.

C. Integration and learning based on mutual exclusivity

1) Integration: The arbitrator integrates the outputs from both modules $\bar{\boldsymbol{r}}_g$ and $\bar{\boldsymbol{r}}_w$ based on the total mutual exclusivity of each module. The integrated configuration segment vector $\bar{\boldsymbol{r}}$ is calculated such as

$$ \bar{\boldsymbol{r}} \quad = \quad \frac{\mu_g \bar{\boldsymbol{r}}_g + \mu_w \bar{\boldsymbol{r}}_w}{\mu_g + \mu_w}, \qquad (3) $$

where μ_g and μ_w are the total mutual exclusiveness, which is calculated as $\mu_g = \max_j\{\bar{r}_{g,j}\}$, and $\mu_w = \max_j\{\bar{r}_{w,j}\}$ respectively. Since each of $\bar{\boldsymbol{r}}_g$ and $\bar{\boldsymbol{r}}_w$ is normalized so that the summation of its elements is equal to 1, they can be larger in the case where only fewer elements have larger values, in other words, there are more mutually exclusive output units. Finally, either element of the integrated configuration segment vector is used as the probabilities of discrete representation of configuration $\boldsymbol{\theta}$ to be selected for JA.

2) Learning: The attentional modules are updated according to the experiences in each trial. A learner cannot always be instructed as to whether the found object is the same as the one the caregiver is looking at. Instead of relying on such detailed instructions, it interprets its experience in an egocentric way: if it finds some object in its focus of attention, it believes that the object is also focused on by the caregiver. It has been shown that it could acquire correct state-action mapping from caregiver's face pattern to joint angles to follow her gaze from her face pattern based on such a belief. Practically the joint angles that occasionally made it find something salient were treated as the backpropagation signal [5] or as rewarded action [6]. Such a belief is considered as valid also in this problem. There is a statistical constraint where the object is more frequently found in the corresponding location in the direction of the caregiver's gaze and it is more frequently matched with the object referred to by the caregiver than at the chance levels under the condition where it observes the gaze and the word label of the caregiver.

Therefore, the connections of the mapping between both the input and output units of which discrete representations are experienced have been strengthened when it finds any object. For the gaze-driven attentional module, the input unit with the maximum element of gaze segment vector and the output unit of the discrete representation for the selected configuration are strengthened. Meanwhile, for the word-driven attentional module, the input unit with the maximum element of word segment vector and the output unit of the discrete representation for the selected object are strengthened. Since we adopted a cumulative representation of correlation in the following simulation, the constant value Δ is added to strengthen the connection weight. At the same time, the other connections to the output units except for the selected one are subtracted Δ_l like lateral inhibition.

III. Simulation of multimodal development

We conducted a computer simulation to show the validity of the μX principle on the developmental process of multimodal joint attention. We first examine the effect of mutual-facilitation in the simplest setting where the caregiver agent

behaves uniformly at random while the learner can perfectly segment the caregiver's gaze and utterances. The segmentation errors are then introduced to argue the effect of the μX principle on the robustness against segmentation errors. Finally, we observe the developmental process of multimodal JA in a more plausible setting for the behavior of the caregiver. After introducing the basic settings for these simulations, the results of the investigation are given in order.

A. Basic settings

Caregiver-infant interaction for instructing word label is simulated in the following experiments. The learner and the caregiver sit across from each other. The table between the caregiver and the learner is divided into 100 locations. Ten objects are randomly selected from 100 candidates and randomly placed on either of these locations. Each object has its own unique word label. In the following experiments, $M_r = 100, M_g = 100, M_w = 100, M_p = 100$.

The caregiver repeats the instruction of a word label of a certain object, that is, she looks at a location where the focused object exists and utters its word label. The procedure in a trial of the interaction is as follows. Ten objects are randomly selected from 100 objects and distributed to locations that are randomly selected from 100 candidates on the table. The caregiver randomly selects one of them, looks at it, and tells its word label. Note that the selection of the caregiver is at random in the first and second experiments while it is done in more careful manner in a way to consequently let the learner proceed "learning from easy mission" [12] in the third experiment. The learner then tries to find out the location the caregiver is referring to by using its attentional mechanism, which is under development. Note that, in the first and second experiments, we also simulate the cases where it uses different attentional mechanisms. The learner finally updates its attentional modules according to the experiences, whether it is successful or not in performing JA. Note that we set σ_g^2 and σ_w^2 to the same values, namely 1.0, except for the simulation to analyze dependency on them in the second experiment. Note also that $\Delta = 1$ and $\Delta_l = 0.01$.

B. Experiment 1: mutual facilitative learning

We first ran 20 sets of 100,000 repetitions of interaction to examine the effect of the μX principle on the learning performance of each module. Fig.4 shows the transitions of the success rate of JA, each of data points and its variance are calculated by counting the success cases in the last 1,000 steps. We can see that the success rates were approaching to 100 % in four cases where the multi-modal attentional modules were used for obtaining experiences. We call the attentional mechanism where the μX principle is applied both for output by Eq.(1) and integration by Eq.(3), *double-μXmultimodal* (blank circles); the one only for each output, *intra-μXmultimodal* (filled triangles); and the one only for integration, *inter-μXmultimodal* (blank triangles). The one without the μX principle is represented by the curve with asterisks. Note that when the intra-μX principle is not applied, output is calculated by regarding that ${}^r\tilde{w}_{g,ij} = {}^r w_{g,ij}$ in Eq.(1) and that when the inter-μX principle is not applied, output is calculated by regarding that $\mu_g = \mu_w$ in Eq.(3). This graph implies that the μX principle for output contributed to the learning speed since the increases of the success rates for double-μX-multimodal and intra-μX-multimodal were much faster than those for inter-μX-multimodal and non-μX-multimodal. This seems because the μX output by Eq.(1) could more strongly bias the learner's experience to be successful than a straightforward output using the observed correlation, once it had experienced successful JA in the similar situation.

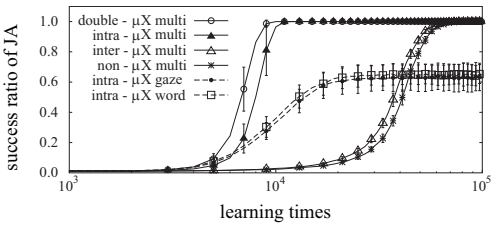
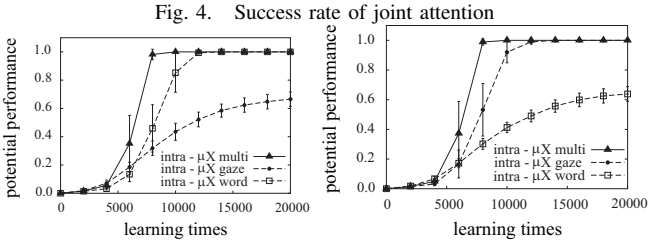


Fig. 4. Success rate of joint attention



(a) Gaze-to-location mapping (b) Word-to-object mapping

Fig. 5. Potential performances of (a) the gaze-to-location mapping and (b) the word-to-object mapping

The success rates of JA did not reach 100 % when only either of the multi-modal attentional modules, i.e., the gaze-driven one (intra-μX-gaze: filled circles) or the word-driven one(intra-μX-word: blank rectangles), was used where the μX principle was implemented only for calculating the output of each module. However, they are useful to illustrate how both modules exhibited mutual facilitation. The potential performance rates of the gaze-driven attentional module (see Fig.5 (a)) and the word-driven one (see Fig.5 (b)) were calculated by letting all possible gaze or word label inputs to either module at each learning step and then counting the cases where it succeeded in outputting the correct location or visual pattern. The potential performance of the gaze-driven module acquired based on intra-μX-gaze (filled circles in Fig.5 (a)) and that of the word-driven module acquired based on intra-μX-word (blank rectangles in Fig. 5 (b)) seem to show that it could not sufficiently improve the gaze or label mapping if it obtained the experiences by using corresponding mapping. On the other hand, there seems to be a facilitative effect on the success rate of the opponent module and a mutually facilitative effect in the case of intra-μX-multimodal (filled triangles in Figs. 5 (a) and (b)). This might suggest that the experiences based on a certain mapping are suitable to modify an opposing mapping but are not suitable to modify itself, and therefore, illustrate

that the μX principle allows the effect of mutual facilitation between constituting each mapping.

C. Experiment 2: robustness against segmentation error

The learner cannot escape from the problem of segmentation, that is how discrete representations can be assigned to the current observed data, especially in the real world situation. If we suppose that the learner must develop not only the mappings but also the mechanism of segmentation, the robustness against the segmentation error would be required for the learning system. Since the current system has multimodal modules, the system had better regard output from a module receiving input with less segmentation error. We expect that the proposed method of integration, that is inter-μX, serves the robustness against such a situation since it determines the contribution weights of modules according to the degree of disambiguation of their outputs.

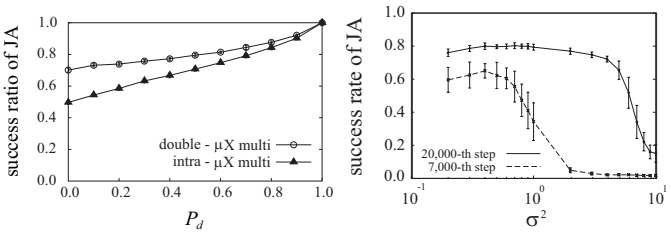
To test this kind of robustness, segmentation errors were introduced into the input layer in the second simulation. We divided the discrete representations for inputs into two groups that involve different degree of the errors. In easy group, we suppose that there are no errors in segmentation. In other words, elements of the input segment vector within this group can be 0 or 1. In difficult group, however, the learner can basically assign correct representation but somehow misassign different ones to the observation. The degree of the correctness of the segmentation for the representation within this group is indicated by p_d called segmentation rate (see Fig.2). When the caregiver's face pattern and/or utterances belong to difficult group, the ambiguous input segment vector is input such as

$$\left[\overbrace{0, \cdots, 0}^{\text{easy group}} , \overbrace{p'_d, \cdots, p'_d, p_d, p'_d, \cdots, p'_d}^{\text{difficult group}} \right]^T, \tag{4}$$

where $p'_d = (1-p_d)/(M_g/2-1)$. Note that halves of the M_g discrete representations of the other's gaze and M_w ones of the word label were set to belong to the difficult group.

We ran the 20 sets of 20,000 repetitions of interactions using two types of attentional modules, where p_d was varied from 0.1 to 1.0. Figure 6 (a) shows the success rates of JA for the last 1,000 interactions in terms of segmentation rate p_d. We can see that the success rates with intra-μX-multimodal (filled triangles) became worth than those with the double-μX-multimodal (blank circles) along with the decrease of the segmentation rate. The results of this simulation and the one in III-C imply that the way of integrating modules based on the μX principle (Eq.(3)) serves the robustness against errors in segmentation while the way of output in each module based on it (Eq.(1)) serves the mutually facilitative effects.

Figure 6 (b) shows the dependency of the learning on the parameters of insensitivity σ_g^2 and σ_w^2 under $p_d = 0.5$ and by using double-μX-multimodal. We can see that the success rates of JA could reach up around 0.8 at the 20,000-th step with large variations (solid curve) while those with limited variations, namely only less than 1.0, could increase at the 7,000-th (broken curve). This implies that we should address a strategy to set or let it adapt to optimal values of these parameters.



(a) Final success rates of JA

(b) Final and intermediate success rates of JA under $p_d = 0.5$

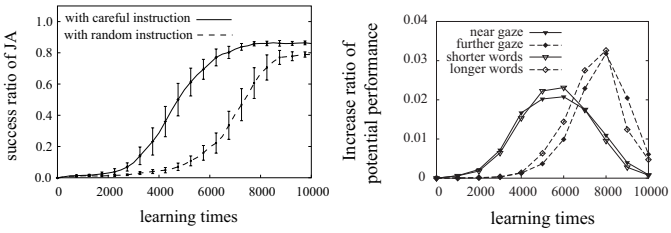
Fig. 6. Success rates under errors in the segmentation process: (a) robustness against the errors and (b) dependency on σ^2

D. Experiment 3: development with caregiver

In the final simulation, we suppose that the caregiver agent behaves more carefully in a way to consequently let the learner proceed by "learning from easy mission" [12]: she starts by referring to closer objects with shorter labels. It seems to resemble human caregivers, and consequently the developmental process more plausibly since she usually assumes that an infant's capabilities of gaze-following and label-comprehension are immature. Furthermore, we suppose that she evaluates the success rate of JA for each location and utterance and based on that extends her choices of objects to be taught to involve more difficult cases where she can still expect to perform JA. Practically, the careful strategy of instruction is implemented as follows. She first tries to select one from those that have labels with which she has succeeded in JA or are located in a position where she has succeeded in JA in two successive trials.

In this simulation, we fixed the attentional mechanism to the double-μX-multimodal and compared the learning progress between cases with different behavior of the caregiver. Figure 7 (a) shows that the success rate of JA with careful instruction (solid curve) grew much faster than one without it (broken curve) where the caregiver selected objects at random. Figure 7 (b) shows the increase ratios of the potential performance rates of the gaze-driven attentional module for gaze direction to a nearby location (filled inverted triangles) and for a farther gaze (filled diamonds) as well as those of the word-driven attentional module for shorter (blank inverted triangles) and longer (blank diamonds) labels. Figure 8 shows the transition of the caregiver's choices: an object with word by which she had succeeded in JA or one at a location where she had succeeded in JA (solid curves) and a closer one with shorter label (broken curves). The potential performance for the difficult groups (broken curves in Fig. 7 (b)) can be seen as synchronized with the occurrence of the difficult choices by the caregiver (solid curve in Fig. 8). They increased after the increase of the potential performances for the easy groups (solid curves in Fig. 7 (b)) more rapidly than them. These

features of the increase of potential performance seem to suggest that there is cross-facilitation: the learning for the difficult group on one module is facilitated by the other module for the easy group.



(a) Success rate of JA (b) Potential performance

Fig. 7. Success rate under careful target choices: (a) comparison to that under random choice and (b) increase ratios of potential performance for the inputs from the easy and difficult groups

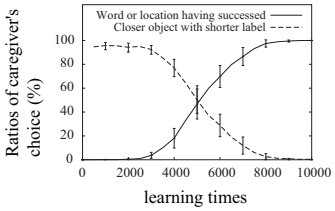


Fig. 8. Transition of the strategy of target choices by a caregiver based on successful experiences of JA (solid curve) or the easiest selection (broken curve)

IV. Discussion and Conclusion

In the series of simulations, we confirmed that the proposed μX principle has the effects of cross-facilitation in learning multi-functions for multimodal joint attention as well as robustness against the errors in segmentation. The simulated developmental process based on the μX principle might be able to reproduce some aspects of an infant's. We can see the gradual extension of the gaze-followable region in the curves of potential performance rates of the gaze-driven attentional module (curves with filled inverted triangles and diamonds in Fig.7 (b)) as observed also in the period of 12 to 18 months of age in infants [1]. Meanwhile, cross-facilitation between the gaze- and word-driven attentional modules in the latter period of the process of extending the gaze-followable region seems to match with the change to refer to the adult's gaze direction for word learning until 18 months of age [2]. Therefore, we might suggest that such a change is synchronized with the potential of the gaze following to the closer region and can be promoted by the careful instructive behavior of the caregiver.

In the simulation of this paper, we assumed that the mechanisms of the segmentation of utterances and gaze direction have been already acquired even though it might involve some errors. Although it seems to follow some observations about the early mechanism of segmentation that 7.5 month-old infants have started detecting sound patterns of words from speech [13] or even that neonates can distinguish whether faces are directed at them or not [14], the mechanism of segmentation seems to keep developing along with mapping process. The mapping based on the μX principle is demonstrated to be not so much affected by the errors in segmentation in III-C. Therefore, it might be a formidable and promising next step to extend the μX principle for improving the mechanism of segmentation along with the development of multimodal mapping. Furthermore, such a mechanism might successfully model the interaction of the developmental processes of segmentation and mapping such as those observed in word learning and phonetic segmentation [15].

References

[1] Butterworth and Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[2] D. Baldwin. Infants' contribution to the achievement of joint reference. *Child Development*, 62:875–890.

[3] E. Bates, P. Dale, and D. Thal. Individual differences and their implications for theories of language development. In Fletcher and MacWhinney, editors, *Handbook of Child Language*, pages 96–151. Oxford: Basil Blackwell, 1995.

[4] M. Asada, K.F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous System*, 37:185–193, 2001.

[5] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15:211–229, 2003.

[6] C. Teuscher and J. Triesch. To care or not to care: Analyzing the caregiver in a computational gaze following framework. In *Proc. of the Third Intl. Conf. on Development and Learning*, 2004.

[7] D. Roy and A. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146, 2002.

[8] C. Yu, L.Smith, Krystal, A. Klein, and R. Shiffrin. Hypothesis testing and associative learning in cross-situational word learning: Are they one and the same? In *Proc. of the 29th Annual Conf. of the Cognitive Science Society*, 2007.

[9] C. Yu, D. Ballard, and R. Aslin. The role of embodied intention in early lexical acquisition. *Cognitive Science*, 2005.

[10] E. Markman and G. Wachtel. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psyhology*, 20:121–157, 1988.

[11] Y. Yoshikawa, K. Hosoda, and M. Asada. Unique association between self-occlusion and double-touching towards binding vision and touch. *Neurocomputing*, 70:2234–2244, 2007. Issues 13-15.

[12] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based reinforcement learning for purposive behavior acquisition. In *Proc. of IEEE International Conference on Robotics and Automation*.

[13] P. Jusczyk and R. Aslin. Infant's detection of the sound patterns of words in fluent speech. *Cognitive Psyhoclogy*, 29:1–23, 1995.

[14] T. Farroni, G. Csibra, and M. Johnson. Eye contact detection in human from birth. *Proc. of National Academy of Science of the United States of America*, 99:9602–9605, 2002.

[15] C. Stager and J. Werker. Infants listen for more phonetic detail in speech perception than in word-learning task. *Nature*, 388:381–382, 1997.