Towards simultaneous categorization and mapping among multimodalities based on subjective consistency

Yuki Sasamoto*, Yuichiro Yoshikawa†, Minoru Asada*
*Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
Email: {yuki.sasamoto, asada}@ams.eng.osaka-u.ac.jp
†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka, 560-8531 Japan
Email: {yoshikawa}@sys.es.osaka-u.ac.jp

***Abstract*—This paper proposes a method for acquiring categories in one modality and mappings between these categories and those in other modalities. Subjective consistency through multimodal mappings is introduced to judge to what extent a perceived signal and inferred ones from other modalities are reliable for categorization and mapping. Based on the proposed method, a simulated infant robot learns categories and mappings by using not only statistics on one perceptual modality but also mappings among the categories in other modalities. The proposed method enables partly simultaneous categorization and mappings.**

I. Introduction

In modeling infant development, it is one of the biggest mysteries how infants acquire categories from their sensorimotor experiences such as its own articulation, the auditory perceptions of the caregiver utterances, and the visual perception of objects. It is supposed that around 3-month-old infants can form some categories from their perceptual information [1], [2]. Statistics of perceptual information are important for such infant categorization (for example, see [3] for audition and [4] for vision). Meanwhile, it is pointed out that categorization is based not simply on the currently perceivable properties of the instances being categorized [5], and more precisely that infants take advantage of both perceptual statistics and theories they hold as they acquire some categories [6]. A recent work on infant object categorization by naming [7] suggests that for 12-month-old infant, "applying the same name to a set of distinct objects (e.g. a duck, raccoon, and dog) highlights the commonalities among these objects and supports the formation of an inclusive category (e.g. animal)." It is also reported a similar effect in 3- and 4-month-old infants [8]. That is, the formulation of categories of one modality depends not only on statistical experiences of its modality but also on prior knowledge of other modalities in early infancy. In order to enable to utilize such prior knowledge, it seems that infants need to acquire mappings of categories among different modalities A question investigated in this study is how we can model the simultaneous developmental processes of categorization and mapping, which depend on each other.

Interdisciplinary approaches are needed since it is difficult to obtain a solution to the question under one single discipline. Among these approaches, a constructive approach seems promising [9]. It aims at understanding human's cognitive development by using physical and virtual robots based on the computational models inspired by neuroscience, cognitive science, and developmental psychology. As one of these studies related to the question above, we have proposed a method of simultaneous development of vocal imitation and lexicon acquisition with a mutually constrained multimodal mapping [10]. The point was introduction of the subjective consistency to judge whether to believe the observed experiences (external input) as a reliable signal for learning or not. A simulated infant robot learned correct mappings among the representations of its caregiver's phonemes, those of its own phonemes, and those of objects even when caregivers do not always give correct examples as real situations. However, they assumed that the input data were already categorized although mapping is actually not always after categorization. Rather, both are simultaneous processes.

Imagine that the robot is going to acquire categories of attended objects and also learn mappings between visual perceptions of these objects and other sensorimotor experiences, such as the motor commands of its own articulation and the acoustic features of sounds heard, which indicates labels of objects. If the robot knows the correct correspondence between objects and those labels, robot can form a category of objects by classifying them according to the commonality of those labels. Also if it has a correct category of objects, the robot can learn the correct mapping between objects and those labels only by associating one object category with its label. However, it is not guaranteed to obtain such a correct association until these categories or mappings mature.

In this study, we extend the previous work [10] by reusing the idea of subjective consistency of multimodal mappings to modify the observed input for making system learn consistent categorization and mappings among multimodalities. In this paper, we report a preliminary computer simulation of

simultaneous categorization and mapping. As a first step, we dealt with an easier learning problem assuming that one of three input layers receives continuous vectors to be categorized while the rest two receive discrete inputs(symbols) as they have finished the categorization process.

The rest of this paper is constructed as follows: First we explain the learning model and introduce the proposed mechanism for categorization through multimodal mappings. We then show the experimental results in computer simulations. Finally, we verify that the proposed mechanism enables to acquire categories based on correct mappings between categories of different modalities .

II. Assumptions

Suppose that a robot and a caregiver take turns in an environment with objects and labels. At each step, the robot looks at an object or vocalize a label. Then the caregiver selects one of three types of behavior: vocalization, showing, and labeling. The robot behavior is assumed to be immature, so the caregiver does not always correctly recognize its utterances or the focus of attention. Therefore, the caregiver is designed to fail to perform such behavior with fixed probabilities that represent not only the robot immaturity but also the tolerance in the caregiver's response. Each type of behaviors is defined as follows:

Vocalization: the caregiver vocalizes the label corresponding to the label uttered by the robot, or utters any labels not corresponding to it. Due to the robot's immaturities for articulation and the caregiver's insensitivities for its utterance, the caregiver is supposed to correctly imitate robot's vocalization with probability p_V.

Showing: the caregiver shows the object whose label uttered by the robot, or shows a different one. Due to the robot's immaturities for articulation and the caregiver inabilities to draw the robot's attention, the probability that the caregiver correctly shows a corresponding object to the robot's utterance is set to p_S.

Labeling (calling): the caregiver utters the label of the object which the robot is looking at, or utters another label. Due to the robot's immaturities for following the caregiver's attention and the caregiver's inabilities to draw the robot's attention, the caregiver is assumed to successfully make the robot see an object and hear a sound label that refers to the object with probability p_L.

In this situation, the robot learns three types of mappings: one is between its own utterance and caregiver one, another is between caregiver's utterance and an object and the other is between an object and its own utterance (see a triangular mappings at the upper part of Fig.1). The robot also forms categories of objects (see the bottom part of Fig.1) in parallel. The robot basically tries to make a connection between two data simultaneously observed on different modalities. However, it is noted that they involve both consistent and inconsistent pairs of data for correspondence among layers. The reason is that the caregiver gives examples of the correct mapping with probabilities p_V for vocalization, p_S for showing and p_L for labeling (hereinafter, these probabilities are collectively called Maternal teaching rate). Otherwise, the caregiver says a label and/or shows an object independently of the robot's utterance or attention.

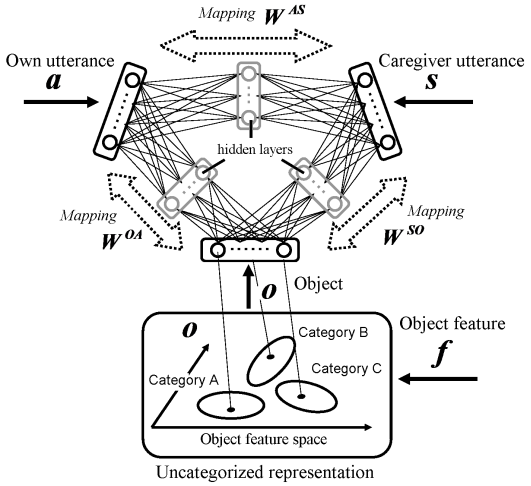


Fig. 1. Multimodal mapping model with an uncategorized representation

III. Multimodal mapping model with uncategorized representation

In this study, we extend our previous model [10] to deal with simultaneous categorization and mapping. Fig.1 shows the multimodal mapping model that consists of three different representations: one corresponds to the robot's articulation space $\boldsymbol{a} \in \Re^{M_i}$, which is assumed to be determined from motor commands of robot's vocalization, another does the sound space $\boldsymbol{s} \in \Re^{M_c}$, which is assumed to be determined from acoustic features of sounds uttered by the caregiver, and the other does the object space $\boldsymbol{o} \in \Re^{N}$ determined from image features of the attending object $\boldsymbol{f} \in \Re^{N_f}$. The robot can obtain one of input vectors on each representation when it vocalizes sounds, when it listens to caregiver's utterances, or when it looks at an object. Each element of these vectors for the robot's vocalization and the caregiver's one is assigned to each node of the corresponding layer. On the other hand, one for the attending object is a two dimensional continuous vector that is a simplified representation of the continuous image feature for object categorization. The object vector is converted to the node activity values using Gaussian mixture models whose likelihoods of kernels are matched with such values. By repeating the interactions, the robot learns three types of the connection weight matrix: $\boldsymbol{W}^{AS}$, $\boldsymbol{W}^{SO}$ and $\boldsymbol{W}^{OA}$.

A. Learning rules of mappings

We employ the mutually associative Boltzmann Machine (hereinafter, BM) [11] as a learning method for mappings. BM is one of the stochastic neural networks and learns the relationship between two observations.

In BM, the output vector $\boldsymbol{y}$ corresponding to observed input vector $\boldsymbol{x}$ is sampled from the following probability distribution:

$$\Pr\left(y_m = 1 | \boldsymbol{W}, \boldsymbol{x}\right) = \frac{1}{1 + \exp\left(-\sum_n w_{nm} x_n\right)}, \quad (1)$$

where y_m is the m-th element in $\boldsymbol{y}$ and x_n is the n-th element in $\boldsymbol{x}$. $\boldsymbol{W}$ is the connection weight matrix between the nodes of input and those of output, and w_{nm} is the element of the n-th row and the m-th column in $\boldsymbol{W}$.

Suppose that the one representation (hereinafter, the first input representation) receives input vector $\boldsymbol{v_1}$ and then another representation (hereinafter, the second input representation) receives other input vector $\boldsymbol{v_2}$. The relationship between these two input vectors are learned as following procedures:

1) Two input vectors $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ are clamped over each representation. Then, the network was allowed to reach equilibrium. Statistics p_{nm} about how often the n-th node in the first representation and the m-th node in the second representation are paired with each other.
2) Only one input vector $\boldsymbol{v_2}$ is clamped over the second representation. Then, statistics p'_{nm} about how often the n-th node in the first representation and the m-th node in the second representation are paired with each other in turn.
3) Connection weight w_{nm} between the n-th node and the m-th node is calculated with two statistics p_{nm} and p'_{nm} as follows:

$$w_{nm} = w_{nm} + \alpha\left(p_{nm} - p'_{nm}\right). \quad (2)$$

At each learning step, the robot calculates the amount of updating not only from the first representation to the second one but also from the second one to the first one. That is, in the above step 2), statistics p'_{nm} is also calculated by clamping the first input vector $\boldsymbol{v_1}$. Connection weight $\boldsymbol{W}$ is updated so that the correspondence of the two representations is constant by duplicating the above.

B. Learning rules of categories

In this study, we employ Gaussian Mixture Models (hereinafter, GMMs) [12] as a categorization method. GMMs are a generative probabilistic model.

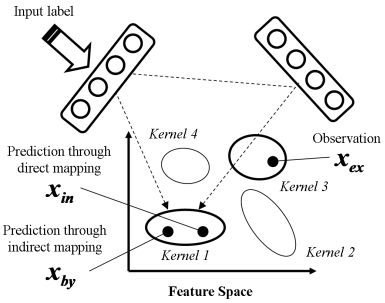
GMMs are a liner superposition of K component Gaussian densities as follows:

$$\Pr(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (3)$$

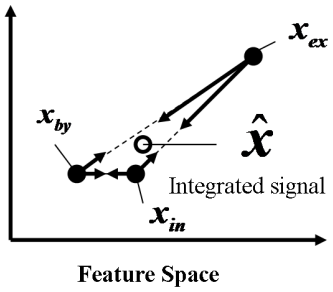
where π_k is a mixture weight of the k-th component. $\boldsymbol{\mu}_k$ is the k-th mean vector and $\boldsymbol{\Sigma}_k$ is the k-th covariance matrix. Gaussian function $\mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is shown as follows:

$$\mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}, \quad (4)$$

where, $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are the D dimensional vectors, $\boldsymbol{\Sigma}$ is $D \times D$ matrix and $|\Sigma|$ is the determinant of $\boldsymbol{\Sigma}$.



(a) First, one observation and two predictions through multimodal mappings are obtained.



(b) Next, these three signals are biased by each other. Then, integrated signal is calculated based on consistency of them.

Fig. 2. Proposed mechanism: Integration based on consistency through multimodal mappings

Suppose that the L sets of D dimensional input vector $\boldsymbol{x}$ was observerd. Each parameter can be estimated by maximum likelihood using EM algorithm in the following steps.

1) **E-step:** Responsibility of the k-th component to the l-th input is calculated as follows:

$$\gamma_{lk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_l | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}_l | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}. \quad (5)$$

2) **M-step:** Then, each parameter is updated as follows:

$$\boldsymbol{\mu}_k = \frac{1}{L_k} \sum_{i=1}^{N} \gamma_{ik} \boldsymbol{x}_i \quad (6)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{L_k} \sum_{i=1}^{N} \gamma_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \quad (7)$$

$$\boldsymbol{\pi}_k = \frac{L_k}{L}, \quad (8)$$

where,

$$L_k = \sum_{i=1}^{L} \gamma_{ik}. \quad (9)$$

Each parameter is updated so that GMMs model the statistical nature of input distribution by duplicating the above.

C. Integration method based on consistency through multimodal mappings

In the previous study, we proposed a learning method for multimodal mappings based on subjective consistency [10]. We extend our proposed method to deal with categorization. As mentioned in the previous section, GMMs can model the statistical nature of input distribution and form some categories as a function of it. However, if a formed category has no correspondence with a category in other modality, it interferes learning of the mapping. For example, imagine that the robot observes some comparable objects which have no consistent label. If the robot forms categories based on only perceptual statistics, it successfully forms the object's category. However, since these objects do not correspond to one label, the category should be wasteful for learning mappings. Moreover, if the robot misunderstands the object as any labeled objects, it can learn wrong mappings. Furthermore, due to the limitation of categorization capacity or the difficulty in predetermining the appropriate number of categories, for example the number of kernels of GMMs that should be assigned for each category, it might be a socially feasible strategy to ignore those that are not labeled by the caregiver instead of trying to categorize everything in the world and fail in forming broad categories. Therefore, in this section, we proposed the method to avoid such a problem by forming any categories depending not only on perceptual statistics but also on predictions through other modalities.

Suppose that the D dimensional external signal $\boldsymbol{x}^{ex}$ and its label are observerd. The activities of the nodes in the layer of the observed label are calculated and propagated to the nodes in the layer of object. Since there are two routes of propagation, one is the route from the input layer to the object layer and the other is one via the other layer, two prediction node signals are generated by the propagation. The prediction node signals from the first route is denoted by $\boldsymbol{a}^{in}$ each of which element is binary value representing the likelihood of each kernel of GMMs while one from the second route is denoted by $\boldsymbol{a}^{by}$. The prediction node signals $\boldsymbol{a}^{in}$ and $\boldsymbol{a}^{by}$ are converted to corresponding continuous vectors $\boldsymbol{x}^{in}$ and $\boldsymbol{x}^{by}$ respectively by using GMMs as follows (see Fig.2(a)):

$$\boldsymbol{x}^n = \frac{1}{\sum_{m\in\{in,by\}} a_m^n} \sum_m^K a_m^n \mathcal{N}'(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \qquad (10)$$

where $\mathcal{N}'(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ represents a manipulation to generate a Gaussian noise whose average is $\boldsymbol{\mu}_m$ and variance is $\boldsymbol{\Sigma}_m$. Then, integrated signal is calculated as follows (see Fig.2(b)):

$$\hat{\boldsymbol{x}} = f(\boldsymbol{x}^{ex}, \boldsymbol{x}^{in}, \boldsymbol{x}^{by}) = \lambda_{ex}\boldsymbol{x}^{ex} + \lambda_{in}\boldsymbol{x}^{in} + \lambda_{by}\boldsymbol{x}^{by}, \quad (11)$$

where λ_n $(n \in ex, in, by)$ represents the subjective consistencies of each learning signal, each of which indicates how it is consistent with others, and is calculated by

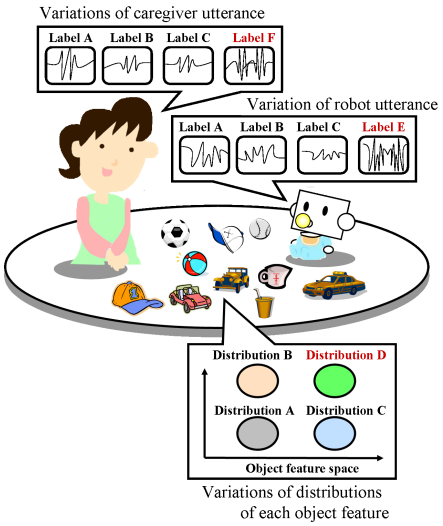


Fig. 3. The environment in simulation

$$\lambda_n = \frac{\exp\left(-e_n/\sigma^2\right)}{\sum_{m\in\{ex,in,by\}} \exp\left(-e_m/\sigma^2\right)}, \qquad (12)$$

where σ is the parameter of sensitivity for the consistencies. e_n represents the consistency of $\boldsymbol{x}^n$ and is calculated with the distances of $\boldsymbol{x}^n$ from other signals such as

$$e_n = \prod_{l\notin n} \|\boldsymbol{x}^n - \boldsymbol{x}^l\|. \qquad (13)$$

In short, the closer to the other two signals, the bigger λ_n is, based on Eqs. (12) and (13).

By creating signals by using not only observed input but also predictions through multimodal mappings, the actual signal is biased to be more feasible for both of categorization and learning mappings. That is, if mappings are mature in part, categories are formed based on predictions calculated from correspondences with labels. If mappings are immature, categories are formed based on statistics of observations and predictions calculated from correspondences with labels. This may enable that categories are formed not only statistical nature of inputs but also mappings between categories of other modalities.

IV. Simulation

To show how the proposed method enables the categorization and learning of multimodal mappings simultaneously, we conducted a computer simulation of caregiver-robot interaction ,as described in Section II.

A. Settings

We assume that the robot can extract labels from the caregiver's utterances and vocalize a label, but it does not know which label uttered by the caregiver corresponds to one uttered by oneself. Let $\boldsymbol{s} \in \Re^M$ and $\boldsymbol{a} \in \Re^M$ be an input that represents which M labels are used for the caregiver and the robot, respectively. For example, if the robot's utterance was /a_i/ that consists of the i-th label, the i-th elements of $\boldsymbol{a}$ was

set to 1, and all other elements were set to 0. We also assume that each object has a distribution in its feature space and robot can observe object features as D dimensional data. $\boldsymbol{f} \in \Re^{N_f}$ is the feature vector of the observed object. Then, $\boldsymbol{o} \in \Re^{N}$ is the node activation vector in the layer for object feature and represents how likely each ID of the kernels of GMMs (i.e., object category) is estimated from the input feature $\boldsymbol{f}$. For example, if it is estimated the k-th object from $\boldsymbol{f}$, the k-th element of $\boldsymbol{o}$ was set to 1, and the other elements were set to 0. However, note that it is unknown for the robot which object feature is generated from which of N kernels of true GMMs and which object corresponds to which labels in the other layers.

In the following experiment, we assume that (1) there are four kinds of objects labeled by A, B, C, and D, (2) the robot can produce four kinds of utterances A, B, C, and E, that are correspond to labels of objects, and (3) the caregiver produces either of four kinds of utterances A, B, C, and F, that are correspond to labels of objects. The above assumptions (1) and (2) mean a situation where the robot does not produce the label of objects that have features of distribution D, but produces another label E that has no correspondence with objects. In other words, M and N_f are set to 4. The assumptions (1) and (3) mean a situation where the caregiver does not pay attention to objects that have features of distribution D, and often produces another label F that does not correspond to any specific object. Furthermore, we suppose the number of the kernel N, that is the maximum number of categories that the robot can form, is 3 which is less than the number of kinds of objects that the robot observes. In such a situation, it might be an adaptive solution to regard the object D as not worth being categorized since it is difficult for the robot to find corresponding labels in other layers.

The following parameters for the learning mechanism were empirically selected for good performance: $\varepsilon = 0.2$, $L = 1000$, $\sigma = 1.0$. We compared the learning performances of the proposed method (hereinafter *integrated*) to those of another method not with integration but with statistics of input (hereinafter *statistical*). The number of the dimension of object features D was set to 2 as a simple case.

B. Results

We ran 10 sets of simulation with 10,000-step interactions for different sets of maternal teaching rates: p_V, p_S, and p_L. These parameters were set to equal each other as p_a and varied from 0.7 to 1.0. Higher p_a means higher percentage of correct pairs of data given by the caregiver to the robot. For example, in the p_a=1.0 case, the caregiver always gave a label corresponding to the robot's attending object. However she randomly selected one label out of A, B, C and E and gave it when the robot saw objects that have features of distribution D. Therefore, it did not always perceive corresponding pairs of labels and objects even if the p_a=1.0. Fig.4 shows the final performance with respect to p_a. Performance was evaluated after 10,000-step interactions by testing 20,000 data generated from distributions which have paring sets (Distribution A, B and C) and checking whether the corresponding labels (Label A, B and C) were recalled. We can see that the performances of the *integrated* (solid curves with filled squires) are apparently higher than those of *statistical* (broken curves with blank triangles).

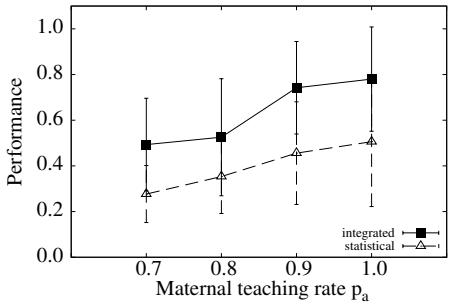
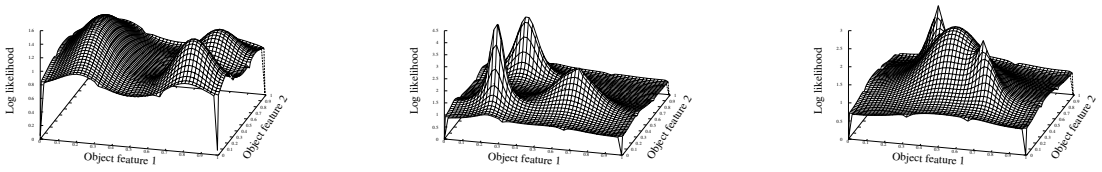


Fig. 4. Final performances in the each condition: solid curves with filled squires shows *integrated* and broken curves with blank triangles shows *statistical*

Fig.5 (a), (b) and (c) show the representative results of categorization through simultaneous categorization and mapping. We can see in *statistical* (Fig.5 (a)) that two distributions corresponding to input distributions C and D respectively are successfully formed. However, one mixed distribution corresponding to input distributions A and B is also formed. This mixed distribution should interfere in learning mappings between label A and Object A or label B and Object B. This causes the low performance of *statistical*. On the other hand, the performance was high in *integrated* since three distributions corresponding to input distributions A, B and C are almost successfully formed (Fig.5 (b)). This shows that integrating a perception and predictions through multimodal mappings based on subjective consistency enables acquiring categories depending not only on statistics of perceptual information in one modality but also on mappings between categories of other modalities. However, if p_a is lower, performance of *integrated* is low ($p_a = 0.7$ in Fig.4) and distributions after learning become more complex(one mixed distribution corresponding to input distributions A and D is formed in Fig.5 (c)). This indicates that due to simultaneous categorization and mapping, it is not guaranteed to converge the correct mapping and categorization if the caregiver almost does not give the correct correspondence to the robot. Other mechanism or bias might be needed in such a situation. Although the use of subjective consistency to modulate simultaneous learning improved the performance compared to one without such modulation.

V. Conclusion and Discussion

In this paper, we proposed a method to bias perception based on the subjective consistency through multimodal mappings to acquire categories and mappings simultaneously. Subjective consistency, which evaluates its closeness to other signals, is used to weight how it contributes to forming categories and learning of the mappings. A computer simulation of caregiver-robot interaction demonstrated that our proposed method enabled simultaneous categorization and mapping. Our proposed method makes the robot to percept observations as more feasible signal to form categories and to learn the



(a) statistical, $p_a = 1.0$ (b) integrated, $p_a = 1.0$ (c) integrated, $p_a = 0.7$

Fig. 5. An example of acquired distribution of categories after learning

mappings. However, if the robot has almost no experiences of the having correct correspondences such as in a low Maternal teaching rate, it is not guaranteed to converge the correct mapping and categorization.

In the current work, the tendency of the caregiver behavior was assumed to be fixed. However, it is indicated that a caregiver facing to her infant shows infant-directed specific behaviors (e.g. *Matherese* [13] and *Mothonese* [14]), and this type of behavior could be scaffolding for the infant to learn several abilities. Also in real situations, the caregivers are very adaptive if they find that infants misunderstand the categorization and mapping, so they directly teach infants by sending correct information. Such scaffolding or direct teaching by caregiver might help infants to highlight different perception among categories or wrong categories already formed. Therefore, we model such a caregiver in simulation and try to verify how such scaffolding helps infant to acquire categories and mappings. We also need to verify to what extent subjective consistency contributes to simultaneous categorization and mapping in such a case.

Furthermore, it has been reported that children exhibit mutually exclusivity bias in language acquisition. That is, they have a tendency to associate novel word with novel object [15]. Such a prior bias might make infant not to confuse or any other words different categories, and therefore facilitate learning mappings. Yoshikawa et al. [16] proposed a method based on mutually exclusivity for associative learning of multi-module, and showed the efficiency of the bias. Trying to fuse these mechanisms is needed.

Several assumptions were introduced in the current experiment (e.g. a constant number of clusters and a two dimensional feature space) to focus on one aspect of difficulties in finding correspondence. Therefore, we need to test whether our proposed method is useful also in more realistic scenarios. For example, by using methods to determine the appropriate number of clusters (e.g. maximum likelihood robust clustering for GMMs [17]), we can release the assumption of a constant number of clusters from the current algorithm. Moreover, several methods for clustering in high dimensional spaces (e.g. high-dimensional data clustering for GMMs [18]) are considered to be applied to cope with higher dimensional input vectors, such as those from sensors and motors of real robots.

ACKNOWLEDGMENT

This study is partially supported by Grants-in-Aid for Scientific Research (Research Project Number: 22220002).

REFERENCES

[1] G.Cameron Marean, Lynne A.Werner, and Patricia K.Kuhl. Vowel categorization by very young infants. *Developmental Psychology*, 28(3):396–405, 1994.

[2] Paul C.Quinn. Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants. *Journal of Ecperimental Child Psycology*, 79:78–94, 2001.

[3] Jenny R.Saffran, Richard N.Aslin, and Elissa L.Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

[4] Natasha Z.Kirkham, Jonathan A.Slemmer, and Scott P.Johnson. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83:B35–B42, 2002.

[5] Linda B.Smith. From knowledge to knowing: Real progress in the study of infant categorization. *Infancy*, 1(1):91–97, 2000.

[6] Sandra R.Waxman and Susan A.Gelman. Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, 13(6):258–263, 2009.

[7] Sandra R.Waxman and Irena Braun. Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. *Cognition*, 95:B59–B68, 2004.

[8] Alissa L.Ferry, Susan J.Hespos, and Sandra R.Waxman. Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, 81(2):472–479, 2010.

[9] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: a survey. *IEEE Trans. on Autonomous Mental Dev.*, 1(1):12–34, 2009.

[10] Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Mutually constrained multimodal mapping for simultaneous development: Modeling vocal imitation and lexicon acquisition. In *Proceedings of the 9th International Conference on Development and Learning*, pages 291–296, 2010.

[11] David H.Ackley, Geoffrey E. Hinton, and Terrence J.Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[12] Christopher M.Bishop. *Pattern Recognition And Machine Learning*, chapter 9: Mixture Models and EM, pages 423–460. Springer-Verlag, 2006.

[13] Fernald Anne and Patricia Kuhl. Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10:279–293, 1987.

[14] Rebecca J.Brand, Dare A.Baldwin, and Leslie A.Ashburn. Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72–83, 2002.

[15] Ellen M.Markman and Gwyn F.Wachtel. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2):121–157, 1988.

[16] Yuichiro Yoshikawa, Tsukasa Nakano, Hiroshi Ishiguro, and Minoru Asada. Multimodal joint attention through cross facilitative learning based on μx principle. In *Proceedings of the 7th International Conference on Development and Learning*, pages 226–231, 2008.

[17] Naoyuki Ichimura. Robust clustering based on a maximum likelihood method for estimation of the suitable mumber of clusters. *IEICE technical report. Pattern recognition and understanding*, 94(241):9–16, 1994.

[18] Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.