Mutually constrained multimodal mapping for simultaneous development: modeling vocal imitation and lexicon acquisition

Yuki Sasamoto*†, Yuichiro Yoshikawa*, Minoru Asada*†
*Asada Synergistic Intelligence Project, ERATO, JST
Email: {yoshikawa, asada}@jeap.org
†Graduate School of Eng., Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
Email: {yuki.sasamoto, asada}@ams.eng.osaka-u.ac.jp

***Abstract*—This paper presents a method of simultaneous development of vocal imitation and lexicon acquisition with a mutually constrained multimodal mapping. A caregiver is basically assumed to give matched pairs for mappings, for example by imitating the learner's voice or labelling an object that it is looking at. However, the tendency cannot be always expected to be reliable. Subjective consistency is introduced to judge whether to believe the observed experiences (external input) as reliable signal for learning. It estimates the value of one layer by combining the values from other layers and external input. Based on the proposed method, a simulated infant robot learns mappings among the representations of its caregiver's phonemes, those of its own phonemes, and those of objects. The proposed mechanism enables correct mappings even when caregivers do not always give correct examples, as real caregivers do not for their infants.**

I. Introduction

Human infants start to comprehend words uttered by adults by eight months and produce their first words by twelve months (the early stage of lexicon acquisition) [1]. They start to mimic the single vowels of adults by eight months as well as consecutive vowels by fourteen months (the early stage of vocal imitation) [2]. Thus, the onset of lexicon acquisition and vocal imitation seems to overlap, and furthermore, one of them might facilitate (or interfere with) the other process. For example, the vocal imitation ability helps infants vocalize unheard words as well as the knowledge of a sound label, and its correspondence to an object helps them imitate the sound label even if it is partially difficult to hear. What kind of mechanisms enable such interaction in simultaneous developmental processes?

In the brain, two different information streams have been studied for their speech perception and production roles. The ventral and dorsal streams underlie mechanisms for sound to meaning mapping (word comprehension) and sound to articulation mapping (vocal imitation), respectively, [3], [4]. Interaction between these streams has been also suggested to exist at the terminal regions of them [4] and underlie the meaning to articulation mapping (word production). Although the literature suggests a mutually constrained network for vocal imitation and lexicon acquisition, it remains unclear how such a structure contributes to the developmental processes of each mapping.

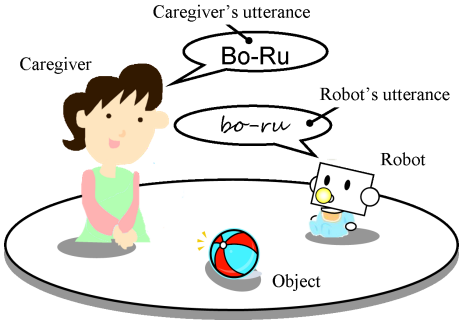


Fig. 1. Typical assumed environment of caregiver-robot interaction

Answering these questions only by developmental psychology or brain science approaches is not easy. Synthetic studies have been considered quite promising approaches for such questions about development mechanisms [5]. In previous work, the process of lexicon acquisition has often been modeled as correlation learning between sound labels and the visual patterns of objects (meaning) referred to as labels [6], [7]. For vocal imitation, correlation learning between infant articulation and phonemes produced by caregivers imitates infant's articulation [8], [9]; the imitative characteristic of caregivers has also been focused on [10]. However, existing synthetic studies on lexicon acquisition and vocal imitation have been done separately. In other words, the interaction in simultaneous developmental processes has not been addressed.

In this study, as in previous work, we assume that a caregiver provides a robot with matched pairs for mappings by three types of behavior: imitating the robot's voice, labeling an object that it is looking at, or displaying an object with its sound label. However, due to the caregiver's arbitrariness and/or the difficulty of correctly inferring its intention, the caregiver's tendency of providing matched pairs for mappings is not always reliable. In this paper, we consider multimodal mappings based on these types of caregiver behavior since

the simultaneous learning of them reduces the difficulty in the learning process of each mapping due to fewer reliabilities in caregiver behavior. For example, imagine that the robot is going to learn mapping between its own articulation and the sound features of the caregiver utterance. Since the caregiver does not always imitate the robot's articulation, a strategy that matches the heard sound with its own articulation often fails. However, if the robot knows other correct mappings, namely, one from its own articulations for sound labels to their meanings and one from meaning to sound features that describe the meaning, it can utilize them to predict the corresponding sound features from its own articulation. Then such a prediction can be utilized to judge whether the heard sound is likely the learning signal for mapping from its previous articulation. However, since such a prediction is not necessarily true until these other mappings mature, a feasible mechanism is needed for a judgment that reflects the learning progress.

For this purpose, we propose a learning method for a mutually constrained multimodal network and introduce an index with subjective consistency to integrate multiple signals fed into a particular layer of the network. The multiple signals include the external signal (observation of a possible match given by the caregiver), a predicted one with a mapping to be learned, and a predicted one with a stream of other mappings. Based on the proposed method, a simulated infant robot learns mappings among the representations of its caregiver's phonemes, those of its own phonemes, and those of objects. The proposed mechanism enables correct mappings even when the caregiver fails to always give correct examples. The rest of this paper is constructed as follows: first we explain the interaction assumptions that a robot must develop and introduce the proposed mechanism. We then show the experimental results in computer simulations. Finally, we verify that the proposed mechanism enables correct mappings even when the caregiver does not always give correct examples as real caregivers do not with their own infants.

II. Assumptions

Suppose that a robot and a caregiver take turns in an environment with objects (see Fig. 1). At each step, the robot looks at either the caregiver or an object and decides whether to say something. 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 modeled to sometimes fail to perform such behavior with fixed probabilities that not only represent the robot immaturity but also the tolerance in the caregiver's response. Each type of behavior is defined as follows:

Vocalization: caregiver imitates the utterances of robot or utters non-imitative words. Due to the robot's immaturities for articulation and the caregiver's insensitivities for its utterance, the caregiver is supposed to correctly imitate with probability p_I.

Showing: caregiver shows an object whose label it utters or 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_C.

Labeling (calling): caregiver shows the robot an object and utters a label that refers to the object. The caregiver selects an object at which it is looking or other objects. 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 it see an object and hear a sound label that refers to the object with probability p_T.

The caregiver selects behavior by consciously or unconsciously taking its attention into account to correctly guide the mapping learning processes. In this study, resembling the likely characteristics of human caregivers who interact with their children, the caregiver's strategy of selecting behavior is assumed to obey the rules below. Note that this strategy is one example of caregiver behavior that not only provides the correct samples of experiments for learning multimodal mapping but also the false ones. When the robot talked while looking at the caregiver, the caregiver always vocalized. When the robot talked while looking at an object, the caregiver always showed that object. When the robot did not talk while looking at the caregiver or an object, the caregiver always said the object label. The caregiver gives examples of correct mappings with probabilities p_I for vocalization, p_C for showing the object on which it focused, and p_T for labeling (hereinafter, these probabilities are called corresponding probabilities). Otherwise, the caregiver says any label and/or shows any object independently of the robot behavior.

The robot does not always get examples for learning correct mappings, which is a situation that resembles actual infants.

III. Mutually constrained multimodal mapping model

Through interaction with a caregiver whose behavior is specified in the previous section, the robot learns mutually constrained multimodal mapping among layers representing its own phonemes $\boldsymbol{a} \in \Re^{M_i}$, those of the caregiver's phonemes $\boldsymbol{s} \in \Re^{M_c}$, and those of objects $\boldsymbol{o} \in \Re^{N}$ (see Fig. 2). It can obtain one of these external input vectors when it vocalizes sounds, when it listens to the caregiver utterances, or when it looks at any object. Each element of these vectors is assigned to each node of the corresponding layer. By repeating the interactions, the robot learns the connection weight matrix between nodes of different two layers, namely, those between its own and the caregiver's phonemes $\boldsymbol{W}^I$ (imitation mapping), those between the caregiver's phonemes and objects $\boldsymbol{W}^L$ (word-listening mapping), and those between the objects and its own phonemes $\boldsymbol{W}^P$ (word-producing mapping).

Suppose that the i-th layer receives input vector $\boldsymbol{x}$ and the j-th layer and then receives another external input vector $\boldsymbol{y}^{ex}$. Given $\boldsymbol{x}$, direct prediction vector $\boldsymbol{y}$ in the j-th layer is

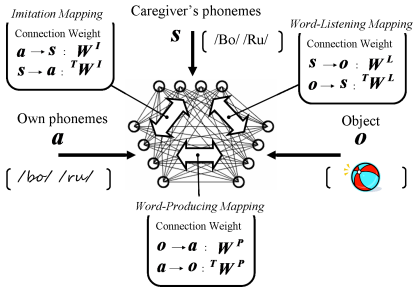


Fig. 2. Mutually constrained multimodal mapping model

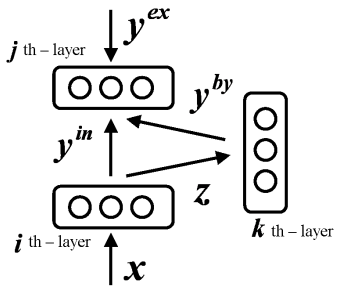


Fig. 3. Notations for learning rules

estimated to predict $\boldsymbol{y}^{ex}$. The m-th element of $\boldsymbol{y}$ 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 the i-th and the j-th layers, and w_{nm} is the element of the n-th row and the m-th column in $\boldsymbol{W}$.

Appropriate values $\boldsymbol{W}$ must be found to correctly estimate the direct prediction vector for vocal imitation or lexicon acquisition. As an example of imitation, the robot is required to learn $\boldsymbol{W}^I$ so that $\Pr\left(\boldsymbol{s}|\boldsymbol{W}, \boldsymbol{x}\right)$ approaches the probability distribution of its own utterance given when it is imitated by the caregiver.

IV. Subjective Integration through Multimodal Representation

In previous works, correlation learning based on external input from caregiver behavior coincident with robot behavior was often considered. For example, when the caregiver imitates the robot's utterance, the robot learns correlation between its own and the caregiver's utterances. If the caregiver always provides the robot with examples of correct mappings through such behaviors as vocalization, showing, or labeling, the robot learns correct mappings using external input from caregivers as learning signals. However, if the caregiver often fails to give such examples, it might learn incorrect mappings.

In mutually constrained multimodal mapping, as in Fig. 2, the values that predict the external input in a certain layer of mappings can be obtained from plural streams through other layers or by directly receiving external input vectors. Therefore, constraining other mappings to be learned by utilizing the predicted values with matured mappings might be feasible to avoid the above problem. However, the learner has to judge which signals are reliable using only accessible variables. In this section, we propose a method of selective integration to create reliable learning signals based on subjective consistency.

Suppose again that $\boldsymbol{x}$ and $\boldsymbol{y}^{ex}$ are external input vectors to the i-th and j-th layers and that $\boldsymbol{y}^{in} (= \boldsymbol{y})$ is the direct prediction vector of $\boldsymbol{y}^{ex}$ from $\boldsymbol{x}$ by direct mapping. Furthermore, suppose another layer labeled by k received from the i-th layer and that outputs indirect (bypassed) prediction vector $\boldsymbol{y}^{by}$ to the j-th layer. Therefore, three vectors $\boldsymbol{y}^{ex}$, $\boldsymbol{y}^{in}$, and $\boldsymbol{y}^{by}$, can be coped with as a possible learning signal (see Fig. 3). Prediction vector $\boldsymbol{y}'$ is calculated as follows:

$$\boldsymbol{y}' = f(\boldsymbol{y}^{ex}, \boldsymbol{y}^{in}, \boldsymbol{y}^{by}) = \lambda_{ex}\boldsymbol{y}^{ex} + \lambda_{in}\boldsymbol{y}^{in} + \lambda_{by}\boldsymbol{y}^{by}, \quad (2)$$

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

$$\lambda_n = \frac{\exp\left(-e_n/\sigma^2\right)}{\sum_{m \in \{ex,in,by\}} \exp\left(-e_m/\sigma^2\right)}, \quad (3)$$

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

$$e_n = \prod_{l, l \notin n} \left\| \boldsymbol{y}^n - \boldsymbol{y}^l \right\|. \quad (4)$$

In short, the closer to the other two signals, the bigger λ_n is, based on Eqs. (3) and (4).

Creating learning signals by using not only external input vectors but also prediction vectors enables caregiver-independent learning for the necessary cases. Furthermore, weighing those signals by the subjective consistencies is expected so that compatible signals can be used as learning signals .

V. Learning Rules of Mappings

We extend the learning rules of the Restricted Boltzmann Machine (RBM) [11], [12] to be mutually associative. RBM, which is a neural network model that consists of both input and hidden layers, is employed as a model of associative memory [13] and prediction [14]. In RBM, the connection weight matrix is updated as follows:

$$\Delta w_{nm} = \varepsilon \left(\langle x_n y_m \rangle_P - \langle \hat{x}_n \hat{y}_m \rangle_{\hat{P}_{\boldsymbol{W}}} \right), \quad (5)$$

where Δw_{nm} is the amount of updating of the connection weight between the n-th element of the input layer and the m-th element of the hidden layer. ε is a learning coefficient, x_n and y_m are the n-th element of input vector $\boldsymbol{x}$ and the m-th element of hidden vector $\boldsymbol{y}$ recalled in response to $\boldsymbol{x}$, respectively. $\hat{x}_n$ is the n-th element of reconstructed input

vector $\hat{\boldsymbol{x}}$ that is recalled using $\boldsymbol{y}$ as input. $\hat{y}_m$ is the m-th element of reconstructed hidden vector $\hat{\boldsymbol{y}}$ recalled in response to $\hat{\boldsymbol{x}}$. P is the probability distribution of input, and $\hat{P}_{\boldsymbol{W}}$ is the probability distribution of input by the model after one step reconstruction. $\langle\cdot\rangle_Q$ denotes an expectation with respect to distribution Q.

Connection weight $\boldsymbol{W}$ is updated so that the correspondence of the input and hidden vectors is constant by duplicating the above process. Therefore, recalling the corresponding patterns of the input and hidden layers is available after learning. In previous works, the signal patterns of the hidden layer were self-organized depending on the input layer. In our study, to recall the corresponding patterns between different external input layers, learning signal $\boldsymbol{y}'$, which is generated based on subjective consistency, is used as an already sampled vector instead of sampling hidden variable $\boldsymbol{y}$. Therefore, the learning rule to update the connection weight between the n-th element of the i-th layer and the m-th element of the j-th layer is modified as follows:

$$\Delta w_{nm} = \varepsilon \left(\langle x_n y'_m \rangle_{PP'} - \langle \hat{x}_n \hat{y}'_m \rangle_{\hat{P}_{\boldsymbol{W}} \hat{P}'_{\boldsymbol{W}}} \right), \qquad (6)$$

where Δw_{nm} is the amount of updating of the connection weight between the n-th element of the i-th layer and the m-th element of the j-th layer. x_n and y'_m are the n-th elements of external input vector $\boldsymbol{x}$ and the m-th element of learning vector $\boldsymbol{y}'$ calculated by subjective integration, respectively. PP' is the joint probability distribution of those signals. $\hat{x}_n$ is the n-th element of reconstructed external input vector $\hat{\boldsymbol{x}}$, which is recalled from $\boldsymbol{y}'$ by Eq. (1). $\hat{y}'_m$ is the m-th element of reconstructed learning vector $\hat{\boldsymbol{y}}'$ calculated by subjective integration under $\hat{\boldsymbol{x}}$. $\hat{P}_{\boldsymbol{W}} \hat{P}'_{\boldsymbol{W}}$ is the joint probability distribution of those signals. Extention from Eq. (5) to Eq. (6) represents that the hidden signals, which were originally self-organized depending on input signals, are biased externally. This is expected to produce mutually associative learning between different external input layers.

At each learning step, the robot calculates the amount of updating not only from the i-th layer to the j-th layer ($\Delta \boldsymbol{W}$) but also from the j-th layer to the i-th layer ($\Delta \boldsymbol{W}'$) using $\boldsymbol{y}^{ex}$ for Eq. (6) instead of $\boldsymbol{x}$. Then it updates $\boldsymbol{W}$ by summing these updates as

$$\boldsymbol{W} = \boldsymbol{W} + \left(\Delta \boldsymbol{W} + \Delta^T \boldsymbol{W}' \right). \qquad (7)$$

Note that the expectations in Eq. (6) are approximated to the sampled or received values in our experiment.

Since we cannot assume that the caregiver always imitates the robot utterances, it is not trivial for the robot to statistically learn correct mappings. In contrast, our proposed method is expected to enable robust learning of mappings against such caregiver error because not only external inputs but also direct/indirect predictions are used for learning. Moreover, since the learning of the three mappings proceeds simultaneously, the effect of mutual constraining is expected to facilitate the learning processes of each mapping: even though it faces a situation where obtaining correct examples for a certain mapping is difficult, if the situation still allows the robot to obtain those for other mappings, it is expected to utilize these easier mappings for learning the difficult one based on subjective consistencies.

VI. Simulation

To show how the proposed method facilitated the learning of mutually constrained multimodal mapping, we conducted a series of computer simulations of caregiver-robot interaction, as described in Section II. In Experiment I, we measured the learning performance under several settings for the corresponding probabilities of three types of caregiver behavior: vocalizing, showing, and labeling. We examined how robustly the proposed method works against lowering corresponding probabilities. To highlight what we call the mutual constraining effect of the proposed method, we measured the learning performance in different settings; fewer examples for correct mappings were given for a specific mapping, and more examples were given for the other two (Experiment II).

A. Common setting

In both experiments, we assume that the robot can extract moras[1] from the caregiver utterances and vocalize any sequence of them but it does not know which caregiver moras correspond to its own. Let $\boldsymbol{s} \in \Re^M$ and $\boldsymbol{a} \in \Re^M$ be an external input that represent which M moras were used for the current caregiver and robot utterances, respectively. For example, if the robot utterance was /a_i a_j/ that consists of the i-th and j-th moras, both the i-th and the j-th elements of $\boldsymbol{a}$ were set to 1, and all other elements were set to 0. External input $\boldsymbol{o} \in \Re^N$ represented which N objects it is looking at. For example, if it looked at the k-th object, the k-th element of $\boldsymbol{o}$ was set to 1, and the other elements were set to 0. Note that $M = 37$ and $N = 39$[2].

One of four types of robot behavior was randomly selected every learning step. Then caregiver behaviors were selected based on the rules described in Section II with parameters specified for each experimental trial. The following parameters for the learning mechanism were empirically selected for good performance: $\varepsilon = 0.2,\ \sigma = 1.0$.

We introduced parameter η to control to what extent the system depends on the proposed method in producing learning signal $\boldsymbol{y}'$, which is determined as follows:

$$\boldsymbol{y}' = (1-\eta)\boldsymbol{y}^{ex} + \eta(\lambda_{ex}\boldsymbol{y}^{ex} + \lambda_{in}\boldsymbol{y}^{in} + \lambda_{by}\boldsymbol{y}^{by}). \qquad (8)$$

Based on this equation, the learning signal is created depending more on the proposed method if η is higher and *vice versa*.

[1] A mora is a phonetic segment with a constant length. For instance, /a/, /ka/, and /bu/ are Japanese moras.

[2] The word labels used in this experiment were selected based on "goo baby" (http://baby.goo.ne.jp) as of February 22th 2009, in which users report when their babies acquire which word labels. We extracted noun words from those reportedly acquired by infants by 18 months.

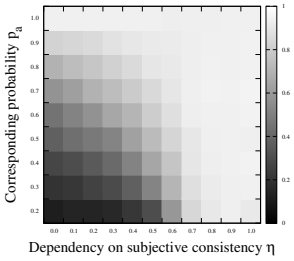


Fig. 4. Average probability of predicting corresponding vectors by acquired mappings until 200,000 steps with respect to dependency on subjective consistency (η) and corresponding probability (p_a)

B. Evaluation

Performance was evaluated in every learning step of the simulation by testing all possible inputs and checking whether the corresponding output vector was closest to the current output sampled from Eq. (1) among all possible outputs. For instance, suppose that the imitation mapping was evaluated and that ${}^i\boldsymbol{s}$ denotes the corresponding vector of the caregiver's sound label to the i-th label to be vocalized ${}^i\boldsymbol{a}(0 < i < 39)$. The imitation mapping is evaluated as the average success ratio of recalling ${}^i\boldsymbol{s}$ from ${}^i\boldsymbol{a}$ and ${}^i\boldsymbol{a}$ from ${}^i\boldsymbol{s}$ among the pairs of correspondence $(0 < i < 39)$. Whether ${}^i\boldsymbol{y}$ is recalled from ${}^i\boldsymbol{x}$ is scored as following:

$$R({}^i\boldsymbol{x}, {}^i\boldsymbol{y}) = \begin{cases} 1 & \text{if } i = \arg\min_j(||{}^i\hat{\boldsymbol{y}}({}^i\boldsymbol{x}) - {}^j\boldsymbol{y}||) \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where ${}^i\hat{\boldsymbol{y}}({}^i\boldsymbol{x})$ is the recalled vector from ${}^i\boldsymbol{x}$. Therefore, the total score S for the imitation mapping is calculated as

$$S = \frac{1}{2}\left(\frac{1}{39}\sum_i R({}^i\boldsymbol{a}, {}^i\boldsymbol{s}) + \frac{1}{39}\sum_i R({}^i\boldsymbol{s}, {}^i\boldsymbol{a})\right). \quad (10)$$

The performances of both word-listening and word-producing mappings are calculated alike.

C. Experiment I: robustness against corresponding probability

We first ran 10 sets of simulations with 200,000-step interactions for different sets of corresponding probabilities: p_I, p_C, and p_T. These parameters were set to equal each other as p_a and varied from 0.2 to 1.0. $p_a = 1.0$ shows the baseline, which is the situation where the robot always get examples of correct mapping from the caregiver. Fig. 4 shows the average final performance of each mapping with respect to η and p_a. Shading means the performance level as of 200,000 steps. Performances with any η is high if p_a is high. However, performances with lower η became worthwhile along with the decrease of p_a, but those with higher η remain high against the decrease of p_a.

Figures 5 (a), (b), and (c) show two transitions of the average performance of each mapping with different parameters of learning mechanism (solid curve for $\eta = 1.0$ and broken one for $\eta = 0.0$) on a condition of less corresponding probability ($p_a = 0.2$). With the former parameter ($\eta = 1.0$), the robot leans depending on the proposed method completely while it does not depend on the proposed method at all with the latter one ($\eta = 0.0$). The performances of $\eta = 1.0$ for all mappings (solid curves) are apparently higher than those of $\eta = 0.0$ (broken curves).

These results show that the method of creating learning signals based on subjective consistency enables robust learning of mappings against corresponding probability. The robot benefits from the fact that the proposed method could adapt to what extent it should rely on the external input that depends on situations. Consistent with this interpretation, subjective consistency for the external input (black squares in Fig. 6) at the final learning period is reduced for cases of less corresponding probability.

D. Experiment II: effect of mutual constraining

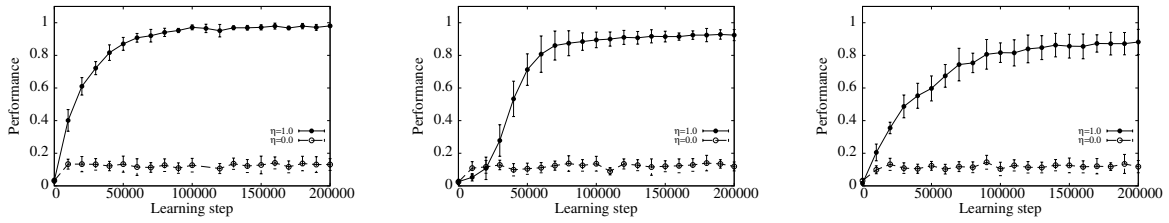
To examine the effect of mutual constraining among different mappings, we ran 10 sets of simulations with 200,000-step interaction for different corresponding probabilities of p_I while p_C and p_T were fixed to 0.4. To see whether relatively matured mappings helped other mappings, that is, facing more difficult learning situations, we set p_I between 0.025 and 0.4. Note that the caregiver showed almost no imitation tendency when p_I was set to 0.025. With such a low value, the probability of giving corresponding moras to those of the robot was less than chance level since $p_I = 0.025(\cong 1/39)$.

Figure 7 shows the average final performances of each mapping with respect to p_I in cases of $\eta = 1.0$ and of $\eta = 0.0$. The performance of $\eta = 0.0$ (black circles) failed to reach high levels even for high corresponding probability. On the other hand, those of $\eta = 1.0$ (white circles) remain high even if p_I almost decreases to the chance level. Although they finally achieved the similar level, the performance of $\eta = 1.0$ at 20,000 steps (asterisks) shows a decrease of learning speed based on the decrease of corresponding probability.

Similar results appeared even where the corresponding probability of other mapping was less than the chance level. Therefore, the proposed method enabled the correct learning of mappings even when a caregiver engaged in such biased behavior to her infant as no imitation, no showing, or no labeling.

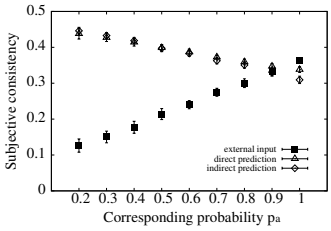
VII. Conclusion and Discussion

In this paper, we proposed a method to combine several sources of a learning signal for mutually constrained multimodal mapping, which is formed by an external input and internal predictions from possible streams of mapping. Each signal's subjective consistency, which evaluates its closeness to other signals, is used to weight how it contributes on creating learning signal through combining with other signals. A series of computer simulations of caregiver-robot interaction demonstrated that our proposed method could model the simultaneous developmental processes of vocal imitation and lexicon acquisition as the learning process of mutually constrained multimodal mapping among representations of the



(a) Imitation Mapping (b) Word-Listening Mapping (c) Word-Producing Mapping

Fig. 5. Average transitions of learning performances: (a) Imitation Mapping, (b) Word-Listening, and (c) Word-Producing with proposed subjective consistency (η = 1.0) and without (η = 0.0)



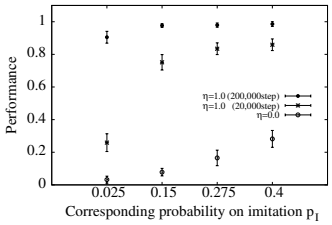


Fig. 6. Average subjective consistencies among three different mappings during final 100 learning steps with respect to corresponding probability p_a

Fig. 7. Performance of imitation mapping with respect to corresponding probability on imitation p_I under $p_C = p_T = 0.4$: those under $\eta = 1.0$ (black circles) and $\eta = 0.0$ (white circles) at 200,000-learning steps as well as those under $\eta = 1.0$ at 20,000-th steps (asterisks).

robot's own phonemes, the caregiver's phonemes, and the objects. Our proposed method makes it possible to successfully ignore external input when the caregiver fails to give examples of correct mapping, which is presumably typical in real caregiver-infant interaction. Note that, given sufficient correct examples for word-listening and word-producing mappings, that is, showing and labeling behavior of the caregiver, the robot learned imitation mapping, even though the caregiver almost completely does not imitate at all.

The proposed method assumes that the robot can receive input vectors representing its sensorimotor experience, such as its own articulation, the auditory perception of the caregiver utterances, and the visual perception of objects. How infants segment and categorize external and internal signals remains a big mystery in modeling infant development. Since the resolution of each representation depends not only on the robot's own modality but also on other modalities, we cope with this issue by synthesizing how such representation can be formed along with the processes of learning mutually constrained multimodal mapping.

Furthermore, in the current work, the tendencies (probabilities) of the caregiver behaviors were assumed to be fixed. However, from the viewpoint of model plausibility for infant development, we should increase the sophistication of the caregiver model by observing caregiver-infant interactions from similar situations in the real world. The more the situation and/or assumptions become realistic, the more human behaviors vary. This might decrease the corresponding probabilities for the robot. Our proposed method is expected to enable correlation learning even in such cases.

REFERENCES

[1] Elizabeth Bates, Philip S. Dale, and Donna Thal. *The Handbook of Child Language*, chapter 4: Individual Differences and their Implications for Theories of Language Development, pages 96–151. Blackwell Publishing, 1995.

[2] Susan S. Jones. Imitation in infancy the development of mimicry. *Psychological Science*, 18(7):593–599, 2007.

[3] Sophie K. Scott and Ingrid S. Johnsrude. The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2):100–107, 2003.

[4] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews*, 8:393–402, 2007.

[5] 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.

[6] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[7] 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.

[8] Frank H. Guenther, Michelle Hampson, and Dave Johnson. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633, 1998.

[9] Hisashi Kanda, Tetsuya Ogata, Kazunori Komatani, and Hiroshi G. Okuno. Vocal imitation using physical vocal tract model. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1846–1851, 2007.

[10] Katsushi Miura, Yuichiro Yoshikawa, and Minoru Asada. Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories. *Advanced Robotics*, 21:1583–1600, 2007.

[11] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[12] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1800, 2006.

[13] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

[14] Graham W. Taylor, Geoffrey E. Hinton, and Sam Roweis. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 19:1370–1378, 2006.