Vowel Acquisition based on an Auto-Mirroring Bias with a Less Imitative Caregiver

Katsushi Miura*, Yuichiro Yoshikawa, Minoru Asada*

Dept. of Adaptive Machine Systems Graduate School of Engineering, Osaka University,
∗ JST ERATO Asada Synergistic Intelligence Project,
yamadaoka 2-1 Suita-City, 565-0871,Japan
katsushi.miura@ams.eng.osaka-u.ac.jp
Dept. of systems Innovation Graduate School of Engineering Science, Osaka University,
machikaneyamachou 1-3 Toyonaka-City, 560-8531, Japan
yoshikawa@sys.es.osaka-u.ac.jp
Dept. of Adaptive Machine Systems Graduate School of Engineering, Osaka University,
∗ JST ERATO Asada Synergistic Intelligence Project,
yamadaoka 2-1 Suita-City, 565-0871, Japan
asada@ams.eng.osaka-u.ac.jp

Abstract

Regardless of interaction with less frequent imitative caregivers, infants can obtain the vowels of the caregivers' mother tongues, by finding the correspondence between their own vowels and the caregivers' ones. This paper proposes a learning method based on the auto-mirroring bias (hereafter, AMB) with a self-evaluation mechanism to find such correspondence. The AMB is the robot's anticipation of being imitated by its caregiver, and has a role of narrowing the candidates for the correspondence. The self-evaluation mechanism biased by the AMB works as outlier (incorrect mapping) rejection expecting that the outliers appear less consistently than the correct mappings do in the interaction. Results from several computer simulations with real sound wave recording from a human experimenter show that the robot could successfully achieve being imitated by the proposed method even if interacting with a caregiver who would seldomly imitate its utterances.

keywords: an Auto-Mirroring Bias, self-evaluation, less imitative

1 Introduction

Humanoid robots have been expected to communicate with humans through the modalities such as voice and gestures which humans utilize. However, due to the difference in physical structure between them, it is difficult for humanoid robots to reproduce the observed humans' actions as they are. This

difference seems to make it also difficult to understand human actions since the cognitive process to reproduce them seems crucial for it, as conjectural from the findings of mirror neuron [1]. On the other hand, human infants seem to successfully solve the same issue regardless of the different articulation structure in the developmental process of language faculty.

Previous studies have demonstrated that a population of computer-simulated agents could self-organize shared vowels through mutual imitation [2, 3, 4]. Fukui et al. [5] showed that the robot can acquire consonant sounds by optimizing the parameters of articulation to minimize the discrepancy between its sound and humans'. However, these studies [2, 3, 4, 5] focused on situations in which all agents can generate sounds in the same region of the acoustic feature space. In other words, they did not consider imitation between dissimilar bodies, which is addressed in this paper.

Kuhl et al. [6] shows that caregiver's social approaches to infant's behaviors are important for its language acquisition. Especially, developmental psychologists have suggested that infant's vowel-like cooing tends to invoke utterances by its mother's imitation [7] and maternal imitation effectively reinforces infant vocalization [20]. However, the mechanisms underlying such facilitation of the infant's language acquisition, namely how an infant learns from a caregiver as well as how a caregiver guides an infant, are still not well understood. As one promising approach to answer these questions, cognitive developmental robotics [9] (hereafter, CDR) has been focused on the mechanisms necessary for such a developmental process.

Yoshikawa et al. [10] constructively showed that imitation by the caregiver in response to infant's vowel vocalization plays an important role in its vowel acquisition. Considering the well-known "perceptual magnet effect" [11] by which a human's perception of phonemes is biased to his/her own category, Miura et al. [12] showed that the utterances of an infant robot could be led to clear vowel sounds through mutual imitation with its caregiver. In addition to this magnet bias, Ishihara et al. [13] have modeled what they call an auto-mirroring bias, which is also necessary for the process of sharing vowels between an infant and its caregiver. It is considered from the fact that in imitative interactions, a caregiver anticipates being imitated by her infant and therefore, perceives the infant's sounds as more closely resembling her precedent utterances. The guiding process of the infant's vowels toward more natural ones for the caregiver was considered in their simulation based on these two biases on the caregiver side. However, they have not considered the effect of these biases on the infant side. It would not be surprising if infants had an auto-mirroring bias. Though it is supposed that the magnet bias (perceptual categorization) has not yet worked since infants do not have vowel categories.

In these studies [10, 12, 13], it was assumed that the caregiver always imitated single vowel-like sounds of the robot but did not utter anything else. However, the mother might not be always motivated to imitate them or might sometimes fail to listen to or to imitate the infants' voice since the infant's vocalization is not yet matured. Louis et al. [14] observed the interaction between mothers and their from 7- to 10-months-old infant and report that the caregiver's vocal imitation of the infant's utterances occurred only in 20 % of cases where the infant utters. To faithfully synthesize the developmental process, we have to consider how a robot can realize being imitated by the caregiver in such low chance

of being imitated.

Infants are supposed to be able to discriminate any vowels before six-months, and their discriminable vowels are gradually tuned to their caregivers' mother tongues after that [15]. This process can be regarded as a mapping process between the caregiver's vowels and some of the infant's vowels through interaction. A challenge of CDR is that an infant robot should find this mapping even if being imitated less frequently in accordance with the data reported by Louis et al. [14]. Here, we propose a learning method based on the auto-mirroring bias (hereafter, AMB) with a self-evaluation mechanism to find the correspondence between the caregiver's utterances and the robot's own ones. The AMB is the robot's anticipation of being imitated by its caregiver, and has a role of narrowing the candidates for the correspondence. The self-evaluation mechanism biased by the AMB works as outlier (incorrect mapping) rejection expecting that the outliers appear less consistently than the correct mappings do in the interaction. Due to the low chance of being imitated, the early mapping is rather poor (chance level), but it gradually improved by the method. Once the robot found the correct mapping, it utters the vowels that can be more easily imitated by the caregiver which helps to obtain more correct mappings. Results from several computer simulations with real sound wave recording from a human experimenter show that the robot could successfully achieve being imitated by the proposed method even if interacting with a caregiver who seldomly imitates its utterances.

2 Basic Idea and Assumptions

Louis et al. [14] observed ten-minutes of unstructured plays, when a mother and her 7- to 10-months-old infant freely play with each other, and analyzed maternal responses appeared in them. They reported that the caregiver responds to them in imitative manners only in 20% of all cases where the responses were made via voice. What makes matters worse is that the physical structures to produce vocal sounds of the mother and her infant are different from each other. Therefore, the caregiver cannot reproduce the infant's voice as it is. In such a severe situation, how can infants find the correspondence between their voices and the caregivers' ?

Here, we propose a computational model to solve this problem focusing on the correspondence of their vowels under the following assumptions.

1. The learner's and the caregiver's articulation regions are different as Fig. 1 illustrates in the sound feature space. Therefore, the caregiver cannot imitate the learner's vowels as they are.

2. The learner can extract the first and second formants, which are well known sound features to discriminate vowels.

3. There are M_v kinds of vowels in the caregivers' language system (for example, M_v=5 in Japanese, i.e., /a/, /i/, /u/, /e/, and /o/). On the other hand, the learner has $m_v(> M_v)$ primitives to utter vowels, but does not know that the caregiver has M_v vowels.

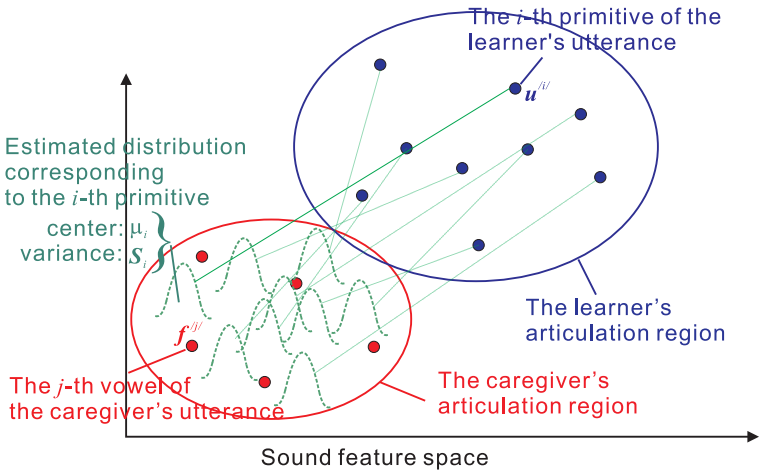


Figure 1: The relationship between the caregiver's vowels and the learner's primitives

4. The caregiver responds to the learner's utterance with an imitative mode (20%) and a non-imitative mode (80%) according to Louis et al.'s analysis [14].

5. In every interaction, the learner initiates to utter and the caregiver certainly responds to the learner's utterance by voice. The learner selects two vowel primitives from its m_v ones (ex. /a/ and /u/) and produce sounds of sequence of them (ex. /au/). Such a form of sound is considered as one of the simplest approximations of continuous utterance in this study.

6. The caregiver has a criterion to decide which vowel among the m_v ones of the learner is easier to imitate with his/her corresponding vowel. If the learner utters an easier one for the caregiver to imitate, the success rate of the imitation increases since the caregiver's imitation sometimes fails even under the imitative mode (20%) if the learner utters a harder one for the caregiver to imitate.

7. The caregiver's utterance included words in addition to the target vowels to be imitated, whether the imitation succeeded or not. For example, when the learner utters *"/au/"* in the the caregiver's imitating mode, the caregiver could say *"Did you say /au/?"* if he/she succeeds in imitating them, while *"You love /eo/!"* if he/she fails ,or in the non-imitative mode, the caregiver could say *"Good boy!"*, *"What do you want?"*, and so on.

The above assumptions 6 and 7 are illustrated in Fig. 2 where the interaction is classified into three cases: one is a non-imitative mode (80%) while the second and the third ones belong to an imitative mode whose probability p^I is 20%. The imitative mode is divided into the successful imitative (the second) mode and non-successful imitative (the third) mode according to whether the caregiver succeeds to imitate or not. The success rate of imitation is parameterized by p_i^S since sometimes the caregiver fails to imitate due to the bad utterance selection by the learner. The task of the learner is to find ways of articulating M_v sounds that the caregiver regards as his/her corresponding ones by rejecting its own $m_v - M_v$ vowels. At the beginning, the leaner has no idea which among m_v vowels corresponds to ones

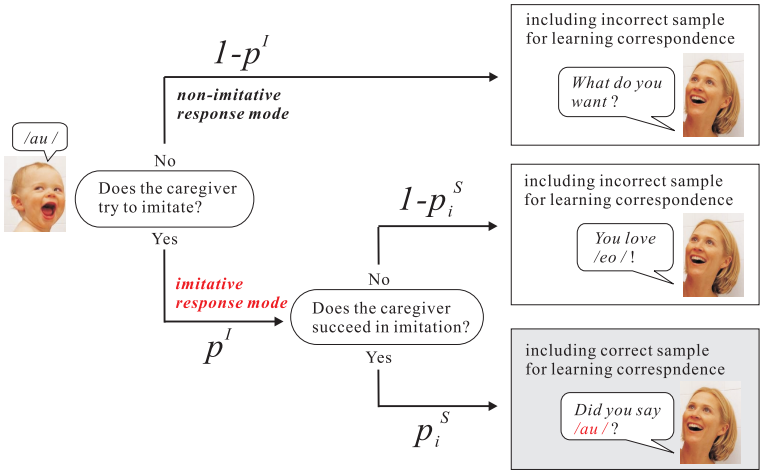


Figure 2: A flow of probabilistic choice of the caregiver's response

of the caregiver. That is, the m_v vowels have equal likelihoods to correspond to one of the caregiver's vowels.

A key idea to accomplish the task is to introduce an AMB (auto-mirroring bias) which biases the likelihoods based on the learner's anticipation of being imitated by its caregiver. Ishihara et al. [13] utilized an AMB on the caregiver's side as an attractor to entrain the imitated utterances by the learner towards the caregiver's own vowels since he/she already knows the correspondence. On the other hand, the main role of an AMB on the learner side is to find the correspondence between the learner's vowels and the caregiver's ones since the learner does not know the correspondence nor how many vowels the caregiver has.

Supposing that the number of vowels of the learner m_v is larger than that of the caregiver M_v, that the caregiver succeeds to imitate if the learner utters the vowels easier for the caregiver to imitate, and that this helps to obtain more correct mappings, the learning will proceed as follows.

1. Distribute m_v models of distributions of the caregiver's sound around the caregiver's articulation region randomly, each of which corresponds to one of the learner's primitive and is used to judge its likelihood whether it hears sounds corresponding to the primitive. Each model is parametrized by the position and the variance of its distribution.

2. Update each models parameters based on relative likelihoods given sound features of the caregiver's response in every interaction. → The AMB puts more weights on the model corresponding to its own utterances supposing to have been imitated and less on the rest.

3. Utter the selected vowels based on the learner's belief how often they are imitated in the past. On the caregiver's side, he/she more often imitates the learner's vowels if they are easier to imitate. → The selection mechanism biased by the AMB works as a learning accelerator.

4. Go back to 2 until the learning converges.

Finally, we expect only M_v models could come to show high likelihoods when hearing the caregiver's utterances while the other $m_v - M_v$ ones remain at low values. The details of the learning mechanism and the experimental results are given in the following sections.

3 Learning mechanism

Fig. 3 shows an overview of the learning mechanism. It consists of classifiers of the heard sound. The number of them is equal to the number of articulation primitives of the learner, that is m_v. Each of these classifiers should learn where (position) and how accurately (variance) the corresponding vowel is located in the caregiver's articulation region. Hereafter, we call these classifiers imitation detectors.

Since the caregiver does not always imitate the learner's voice, the learner should notice being imitated by herself. In this section, we propose a self-organizing method of the learning imitation detector not only based on calculating a likelihood of each classifier but also based on biasing itself by what we call the auto-mirroring bias (AMB in Fig. 3). We denote the sound feature of a certain time window ΔT, consisting of the first and second formants by $\boldsymbol{f}$. The number of the primitives or classifiers is set to m_v. The sound produced by the i-th primitive is denoted by $/v_i/, (i = 1, 2, \cdots, m_v)$.

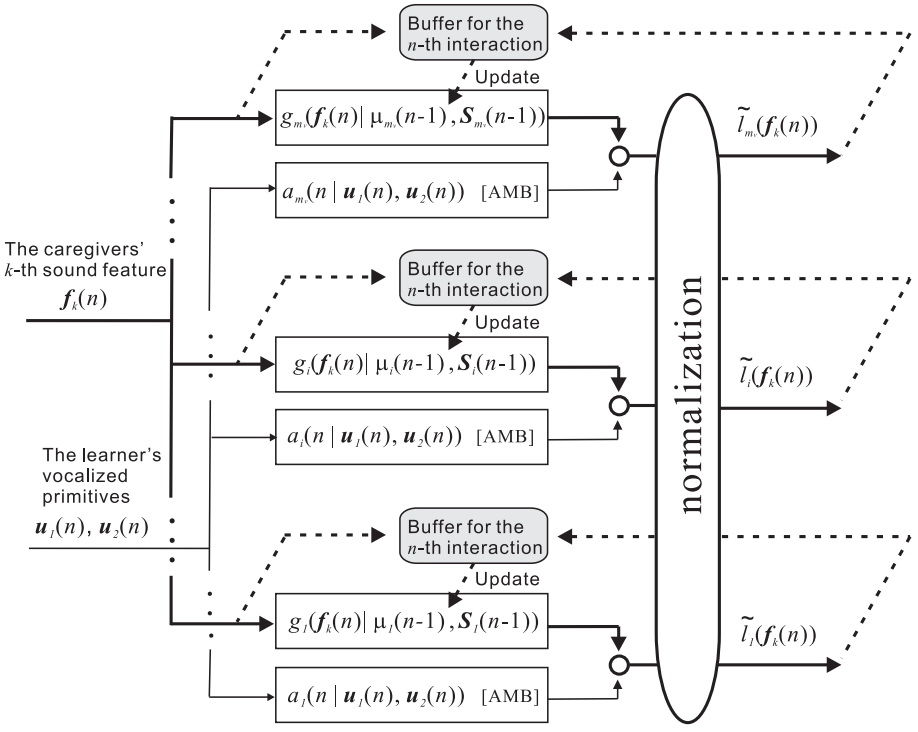


Figure 3: An overview of the learning mechanism

3.1 An imitation detector

We utilize a two dimensional Gaussian function in the space of the first and the second formants as a classifier that models a distribution of the careigver's sound observed when the caregiver imitates specific vowel primitives of the learner. The Gaussian function for $/v_i/$ is denoted by $g_i, (i = 1, 2, \cdots, m_v)$ whose parameters are $\boldsymbol{\mu}_i$ and $\boldsymbol{S}_i$, representing its center and the covariance matrix. Since g_i would be used for detecting the sound feature when the caregiver imitates the i-th primitive of the learner, we call g_i an imitation detector for the i-th primitive.

The sound analysis is first applied to the caregivers' utterance to obtain sound features within each moving time window ΔT. The learner obtains a sequence of the sound features in every interaction. We denote the k-th sound feature in the n-th interaction by $\boldsymbol{f}_k(n)$ (see Fig. 4).

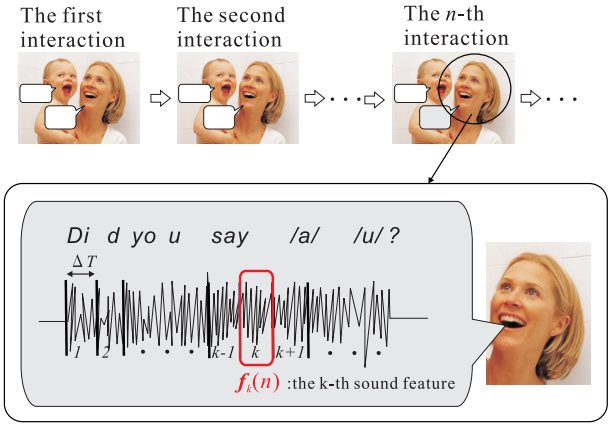


Figure 4: The k-th sound feature $\boldsymbol{f}_k(n)$ in the n-th interaction

To update the parameters of g_i, the learner has to extract the segments of the sound in the timing when the caregiver imitates the learner's utterance with the i-th primitive. For $\boldsymbol{\mu}_i(n)$, it is updated by calculating a weighting average of the sound features $\bar{\boldsymbol{f}}^i(n)$ according to $l_i(\boldsymbol{f}_k(n))$ (described in the next section), that is the likelihood of g_i given $\boldsymbol{f}_k(n)$, as follows

$$\bar{\boldsymbol{f}}^i(n) = \frac{\sum_k l_i(\boldsymbol{f}_k(n))\boldsymbol{f}_k(n)}{\sum_k l_i(\boldsymbol{f}_k(n))} \quad . \tag{1}$$

At the n-th step, the parameter of its center $\boldsymbol{\mu}_i(n)$ is then updated by moving average among the history of $\bar{\boldsymbol{f}}^i(n)$ such as

$$\boldsymbol{\mu}_i(n) = (1-\eta)\boldsymbol{\mu}_i(n-1) + \eta\bar{\boldsymbol{f}}^i(n) \quad , \tag{2}$$

where η is a forgetting factor and set to 0.99 in this study. At the same step, the variance $\boldsymbol{S}_i(n)$ is updated as follows:

$$\bar{\boldsymbol{S}}^i(n) = \frac{\sum_k l_i(\boldsymbol{f}_k(n))(\boldsymbol{\mu}_i(n) - \boldsymbol{f}_k(n))(\boldsymbol{\mu}_i(n) - \boldsymbol{f}_k(n))^T}{\sum_k l_i(\boldsymbol{f}_k(n))} \quad \text{, and} \tag{3}$$

$$\boldsymbol{S}_i(n) = (1-\eta)\boldsymbol{S}_i(n-1) + \eta\bar{\boldsymbol{S}}^i(n) \quad . \tag{4}$$

3.2 Imitative likelihood l_i

The imitative likelihood of the i-th imitation detector given $\boldsymbol{f}_k(n)$ is calculated as the relative value of likelihoods of all other imitation detectors as follows:

$$l_i(\boldsymbol{f}_k(n)) = \frac{g_i(\boldsymbol{f}_k(n)) + \epsilon(n)}{\sum_{j=1}^{m_v}\{g_j(\boldsymbol{f}_k(n)) + \epsilon(n)\}} \quad , \tag{5}$$

where $\epsilon(n)$ is a positive small value for the stability of calculation and $g_i(\boldsymbol{f}_k(n))$ is the absolute likelihood of the i-th detector that is calculated by using a Gaussian function with the estimated parameters $\boldsymbol{\mu}_i(n)$, and $\boldsymbol{S}_i(n)$, as follows:

$$g_i(\boldsymbol{f}_k(n)) = \frac{1}{\sqrt{2\pi|\boldsymbol{S}_i(n)|}} \exp\{-\frac{(\boldsymbol{f}_k(n) - \boldsymbol{\mu}_i(n))^T \boldsymbol{S}_i(n)^{-1}(\boldsymbol{f}_k(n) - \boldsymbol{\mu}_i(n))}{2}\} \quad . \tag{6}$$

To stabilize the calculation in the beginning of learning, the parameter $\epsilon(n)$ is added to both the numerator and the denominator in (5). Along with the learning progress, $\epsilon(n)$ is scheduled to gradually decrease by the following equation:

$$\epsilon(n) = \epsilon_0 + \frac{\alpha}{1 + \exp\{\beta \cdot (n - \gamma)\}} \quad , \tag{7}$$

where α, β, γ, and ϵ_0 are parameters to control $\epsilon(n)$ and set to 0.999, 0.005, 2000, and 0.001.

3.3 The auto-mirroring bias (AMB)

We suppose the effect of the AMB is not only to attract to the learner's own perception as assumed by Ishihara et al. [13], but also to bias towards a believe in the caregiver's imitation: to learn the correspondence between the own vocalized vowel and the caregiver's response. Here, we examine the effect of the AMB to believe and learn the caregiver's response as the imitation corresponding to the robot's own vocalized vowel whether the caregiver imitates or not.

We introduce another variable $a_i(n)$ into the calculation of the imitative likelihood to consider the AMB, which varies according to whether the i-th primitive is used in the n-th interaction such as

$$a_i(n) = \max\{\bar{a}_i(\boldsymbol{u}_1(n)), \bar{a}_i(\boldsymbol{u}_2(n))\} \quad , \tag{8}$$

where $\boldsymbol{u}_1(n)$ and $\boldsymbol{u}_2(n)$ are formants that are produced by the learner in the first and second vowel selected in the n-th interaction while $\bar{a}_i(\boldsymbol{u})$ is the indicator of the AMB for the i-th primitive when it produces $\boldsymbol{u}$. It is calculated as follows:

$$\bar{a}_i(\boldsymbol{u}) = \exp\{-\frac{\|\boldsymbol{u} - \boldsymbol{u}^{/v_i/}\|^2}{2\sigma^2}\} \quad , \tag{9}$$

where $\boldsymbol{u}^{/v_i/}$ is the target formant that the learner should hear when it vocalizes the i-th primitive while $\sigma(=50)$ is a parameter to control the sensitivity of the AMB.

The equation to calculate the imitative likelihood is replaced as follows:

$$\tilde{l}_i(\boldsymbol{f}_k(n)) = \frac{a_i(n)g_i(\boldsymbol{f}_k(n)) + \epsilon(n)}{\sum_{j=1}^{m_v}\{a_j(n)g_j(\boldsymbol{f}_k(n)) + \epsilon(n)\}} \quad , \tag{10}$$

where $\tilde{l}_i$ is the modified imitative likelihood including the AMB. Updating parameters in (1) and (3) is proceeded with $\tilde{l}_i$ instead of l_i.

3.4 The choice of the robot's primitive

In every interaction, the robot chooses two primitives to utter. The choice is done based on to what extent it expects the caregiver to imitate each primitive. The degree of such an expectation of the i-th primitive $/v_i/$ is denoted by e_i.

3.4.1 Degree of expectation e_i

The degree of expectation for $/v_i/$ is calculated from the history of likelihood l_i of the i-th primitive. It is considered that $l_i(\boldsymbol{f}_k(n))$ becomes biggest at the imitative sound among all segments in the caregivers' utterance in the n-th interaction, i.e. $\bar{e}^i(n)$ in (11) is considered to roughly indicate whether the sound produced by the i-th primitive is imitated by the caregiver in the n-th interaction. The degree of expectation denoted by e_i is defined as the average value of the $\bar{e}^i(n)$ among the history of interaction as follows:

$$\bar{e}^i(n) = \max_k \; l_i(\boldsymbol{f}_k(n)) \quad \text{, and} \tag{11}$$

$$e_i(n) = \frac{n-1}{n}e_i(n-1) + \frac{1}{n}\bar{e}^i(n) \quad . \tag{12}$$

3.4.2 Selection probability $q_i(n)$

We define selection probability $q_i(n)$ to refer to the probability of the i-th primitive to be selected to produce the learner's utterance in the n-th interaction. $q_i(n)$ is calculated by the following equation

$$q_i(n) = \frac{e_i^2}{\sum_{j=1}^{m_v} e_j^2} \quad . \tag{13}$$

4 Experiments

We conducted computer simulations of a semi-realistic human-robot interaction using real sound waves recorded from a person. We simulated a mother-infant interaction as observed in the literature [14], which involves the difficulty of finding correspondence of vowels between them. It was shown that the simulated learner could find corresponding vowel primitives to the ones of the caregiver from less frequently imitative interaction with the caregiver by considering the infant-side aspect of the AMB.

4.1 Setup

It was supposed that, in every interaction, a learner first selects two primitives from m_v primitives of articulation and a caregiver agent then responds to the learner's utterance according to the parameters of p^I and p_i^S as described in section 2. Ten runs of computer simulations of 10,000 times interaction were conducted to analyze tendencies of the found correspondence as the parameters of Gaussian functions $g_i(\boldsymbol{\mu}_i(n), \boldsymbol{S}_i(n)), (i = 1, 2, \cdots, m_v)$.

The number of articulation primitives of the learner (i.e., m_v) is 100. We show the first and the second formants of the learner's clearest vowels $\boldsymbol{u}^{/v_i/}(i = 1, \cdots, 5)$, each of which corresponds to /a/, /i/, /u/, /e/, and /o/, in Table 1. The other primitives were assumed to be distributed around the clearest vowels. They are assumed to be categorized into five categories each of which corresponded to one of the Japanese vowels. Since these primitives in one category are assumed to have different clarity for the caregiver, different values were assigned to each p_i^S. The parameter p_i^S is set to be higher according to how closer a sound the i-th primitive could generate to the clearest ones in the same category. The 100 primitives are assumed to be distributed so that each category has twenty levels of clearity sounds as shown in Table 2. Here, the first five primitives $/v_1/, \cdots, /v_5/$ were considered to be the clearest corresponding vowels /a/, /i/, /u/, /e/, and /o/ for the caregiver.

Table 1: The first and the second formants of the learner's clearest vowels (Mel scales)

	1st formant	2nd formant
$\boldsymbol{u}^{/v_1/}$	1243.5	1609.7
$\boldsymbol{u}^{/v_2/}$	770.3	2146.8
$\boldsymbol{u}^{/v_3/}$	812.1	1538.4
$\boldsymbol{u}^{/v_4/}$	981.6	1981.6
$\boldsymbol{u}^{/v_5/}$	941.8	1463.0

Table 2: Vowel category and p_i^S assigned to each articulation primitive $/v_i/, (i = 1, 2, \cdots, 15)$

	/a/	/i/	/u/	/e/	/o/
$p_i^S = 1.0$	$/v_1/$	$/v_2/$	$/v_3/$	$/v_4/$	$/v_5/$
$p_i^S = 0.95$	$/v_6/$	$/v_7/$	$/v_8/$	$/v_9/$	$/v_{10}/$
$p_i^S = 0.9$	$/v_{11}/$	$/v_{12}/$	$/v_{13}/$	$/v_{14}/$	$/v_{15}/$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p_i^S = 0.05$	$/v_{96}/$	$/v_{97}/$	$/v_{98}/$	$/v_{99}/$	$/v_{100}/$

A person's utterances were recorded and used as the voice of the caregiver in the simulation. The sampling frequency was 11025[Hz] and the sampling window for analyzing acoustic features was 256

samples of the segmented caregiver's responses. Moreover, data with lower levels of sound power was cut as noise. The first and second formants were analyzed for each sampling window in the mel scales. Conversion of the sound feature in the Hz scale s_h into the mel scales s_m was performed in the following equation

$$s_m = \frac{1000}{\log_{10} 2} \log_{10}(\frac{s_h}{1000} + 1) \quad . \tag{14}$$

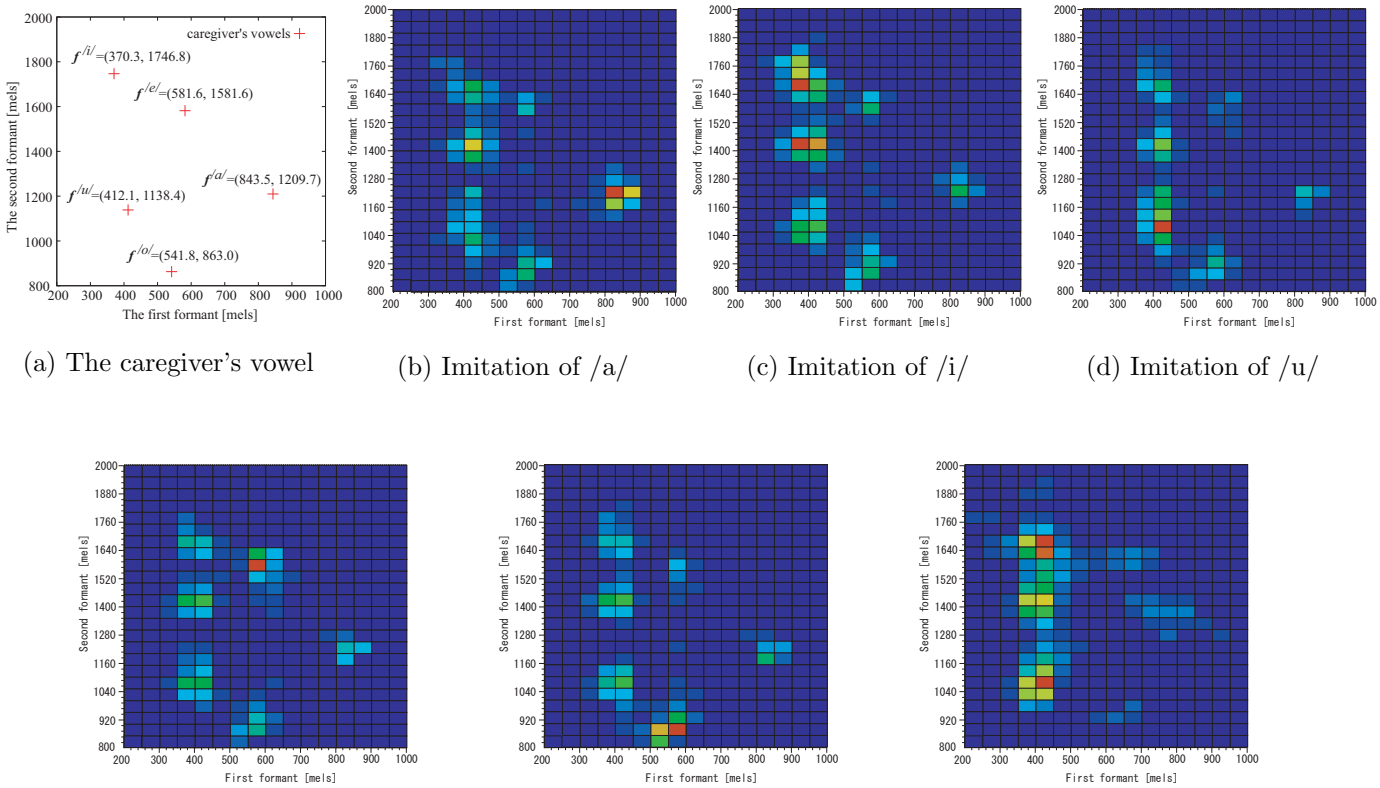
In the imitative response mode, the caregiver replied in either of the following five types of utterances: "/X_iX_j/", "/X_iX_j/ KANA?" ("Did you say /X_iX_j/?"), "/X_iX_j/ NOKOTO?" ("You mean /X_iX_j/?"), "KOREHA /X_iX_j/ DESU" ("This is /X_iX_j/."), "SOUSOU, /X_iX_j/!" ("Yes, /X_iX_j/"). Here, /X_iX_j/ $(i, j = 1, \cdots, 5)$ indicates a sequence of two vowels selected from the five Japanese ones. Therefore, 125 combinations are possible, which are combinations of different sentence types and uttered vowels. For the non-imitative response mode, the other 125 utterances were recorded. We asked the person to utter the following sentences 25 times: "DOUSITANO?" ("What's wrong?"), "ONAKASUITA?" ("Are you hungry?"), "NANINANI?" ("What did you say?"), "MOUIKKAIITTE" ("Please say it one more time."), and "YOKUDEKITANE" ("Good boy!."). After the learning process has successfully proceeded, the learner is expected to be able to exclude such parts of non-imitative sentences according to the value of $\tilde{l}_i$.

We show the average values of the first and the second formants of each vowel which appeared in the total 250 utterances of the caregiver in Fig. 5 (a). Histograms of formants in the imitative utterance of the learner including clearest vowels, that is /v_1/ (Fig. 5 (b)), /v_2/ (Fig. 5 (c)), /v_3/ (Fig. 5 (d)), /v_4/ (Fig. 5 (e)), /v_5/ (Fig. 5 (f)), which correspond to /a/, /i/, /u/, /e/, and /o/, respectively are shown in Fig. 5. In these figures, the horizontal and vertical axes indicate the first and second formants. The number of formants in the caregiver's utterances is represented by color intensity from blue (less frequent) to red (more frequent). Fig. 5 (g) shows a similar histogram in the non-imitative mode.

As shown in figures from Fig. 5 (a) to (f), the most frequently vocalized sounds in the caregiver's imitative response mode are the corresponding vowels to the learner's. Therefore, the learner would have many chances to learn a correct vowel mapping, when the caregiver succeeded in imitating the leaner's utterances. On the other hand, as shown in Fig. 5 (g), the caregiver's responses in the non-imitative mode are widely distributed near the vowel /i/ or /u/.

Therefore, the robot frequently learns the relationship between its own vowels and the caregiver's vowel /i/ or /u/, and the center of Gaussian function $\boldsymbol{\mu}_i(n)$ also moves to the area near the vowel /i/ and /u/. From Fig. 5, it is inferred that the robot cannot learn the correct relationship between own vowels and the caregiver's vowels if the caregiver does not imitate at some frequency. Therefore, the learner would be incorrectly biased to associate all its vowels to the caregiver's vowels of /i/ and /u/ while its vowels to be associated with /i/ or /u/ would be also biased to others.

We ran learning simulations with five different parameters of p^I, namely 0.01, 0.05, 0.1, 0.2, 0.5, and 1.0. The initial values of the parameter $\boldsymbol{\mu}_i(n)$ of the i-th imitation detector was selected at random so



(a) The caregiver's vowel (b) Imitation of /a/ (c) Imitation of /i/ (d) Imitation of /u/

(e) Imitation of /e/ (f) Imitation of /o/ (g) non-imitative mode

Figure 5: The caregiver's vowels and the tendency of the caregiver's responses

that the Gaussian center indicated by them was distributed inside of 200 mel across a circle centered at the centroid of the caregiver's vowels. The initial values of the parameter $\boldsymbol{S}_i(n)$ were calculated from the segments in the possible 250 patterns of the caregiver's utterance.

4.2 Results

Fig. 6 shows the average profiles of the learning progress among ten trials: how the imitation detectors for the clearest primitive approach the desired values (blue curve) and how the clearest primitives become to be selected more frequently through learning (red curve). The former was evaluated as average discrepancy between the estimated centers of the imitation detectors and the corresponding caregiver's vowels. We can see that it converged to about 50 [mels], which is much less than the discrepancy between the caregiver's vowels so that it can distinguish them by the estimated detectors. On the other hand, the latter was evaluated by summing q_i for the clearest primitive of each category. We can see that the averaged selection probability of clearer primitives became higher than less clear ones. The drastic decrease of the estimation error and increase of the selection probability for the clearest primitives appears around the 3,000-th step. It is considered that such a synchrony is caused by the change in $\epsilon(n)$ that corresponds to the change of the learner's confidence in the self-evaluation.

The details of the learning progress can be seen in an example transition of the learning progress where $p^I = 0.2$ (Fig. 7): (a) at the beginning, (b) at the 2,000-th interaction, (c) at the 3,000-th

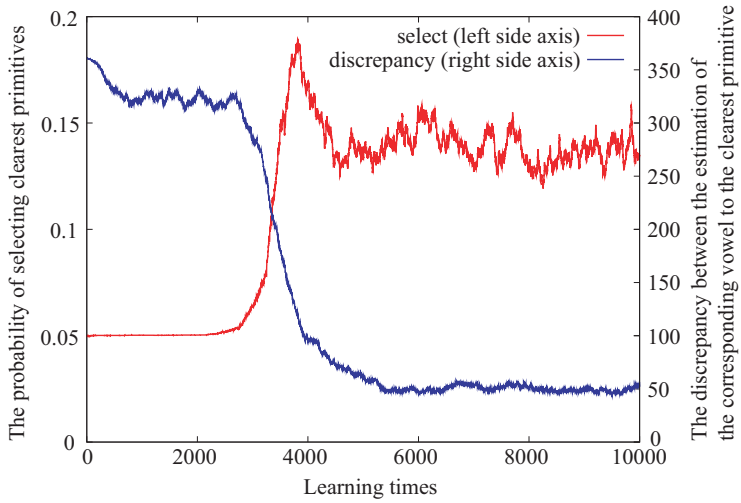


Figure 6: The transitions of sum of selection probability q_i of the clearest primitives and the average discrepancy between the estimation of the corresponding vowels to the clearest primitive of each category under the condition with the imitation probability $p^I = 0.2$

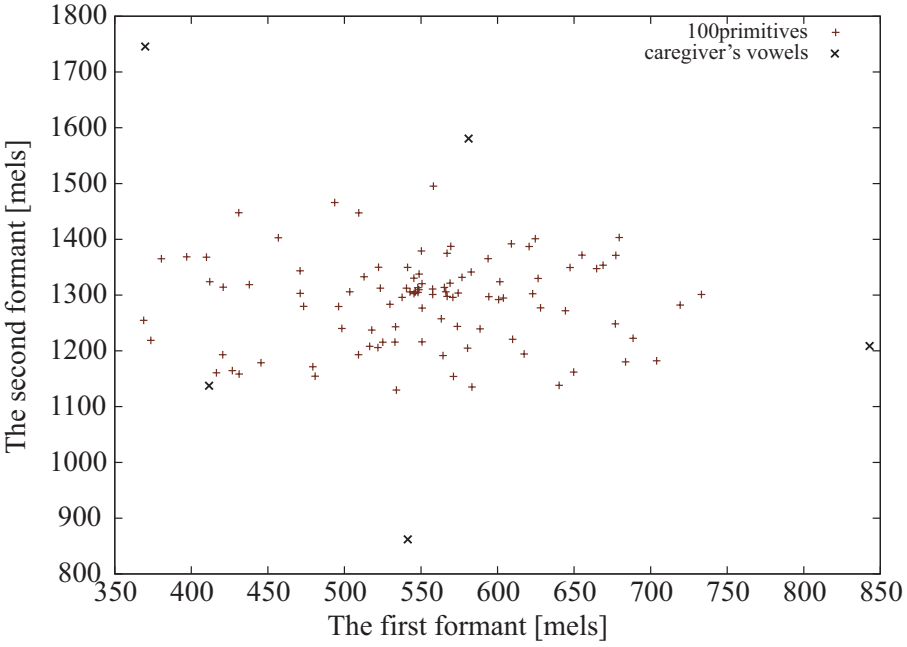
interaction, (d) at the 4,000-th interaction, and (e) at the 10,000-th interaction. In these figures, transitions of the estimated centers of the Gaussian functions in the imitation detectors are shown as brown dots while trajectories of the clearest primitives of each vowel category are shown with solid lines of different colors, that is /a/ (red), /i/ (green), /u/ (blue), /e/ (purple), and /o/ (sky blue). As shown in Figs. 7, starting from the initial placements, the estimated centers of the clearest primitives first converge to the center and then approach the corresponding caregiver's vowels, which are indicated by character the "+". The reason of the initial convergence is that the value of $\epsilon(n)$ in equation (10) is relatively larger than $g_i(\boldsymbol{f}_k(n))$ at the beginning of learning. This makes $\tilde{l}_i(\boldsymbol{f}_k(n))$, which is used as weighting factor of the data constant independent of inputs and therefore the estimated averages for all primitives are converged to the some center of all inputs.

Next, we tested whether the learner became able to imitate the caregiver's vowels. We input the caregiver's vowels to the imitation detectors and let the learner select one primitive i^* according to the likelihood using equation (5):

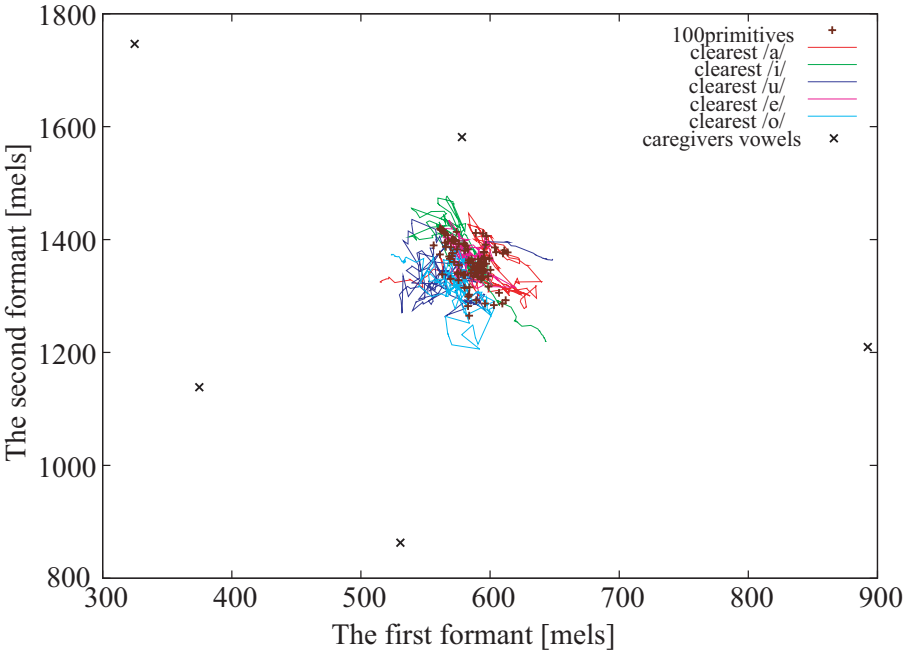
$$i^* = \arg\max_i l_i(\boldsymbol{f}^{/k/}) \quad , \tag{15}$$

where /k/ is one of the caregiver's vowels, that is /a/, /i/, /u/, /e/, and /o/. The values of $\boldsymbol{f}^{/k/}$ are shown in Fig. 5 (a).

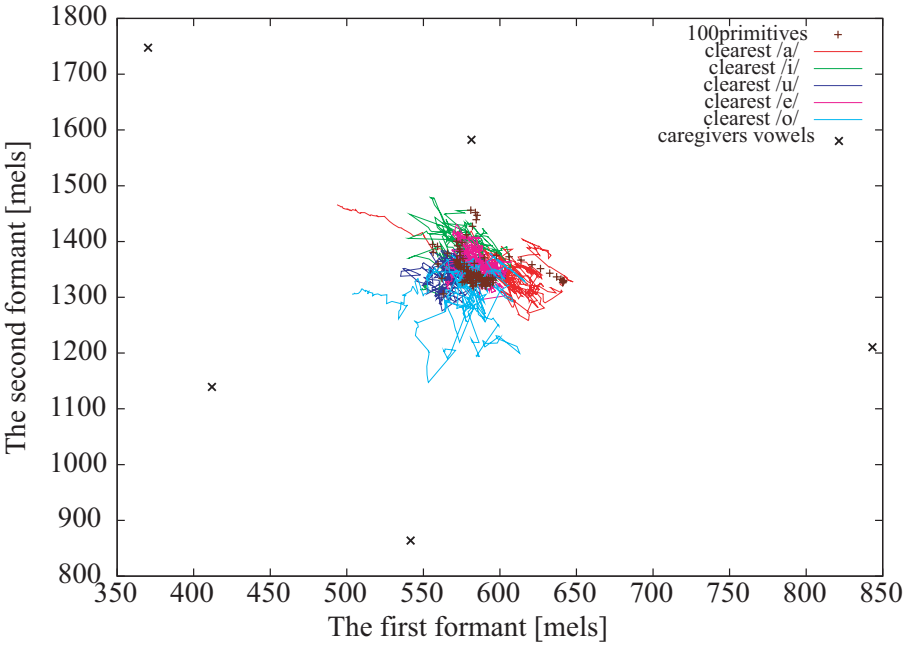
The success rate of the learner's imitation was then evaluated according to the probability that the caregiver would succeed in imitating it again. It can be calculated by averaging p_i^S of the selected



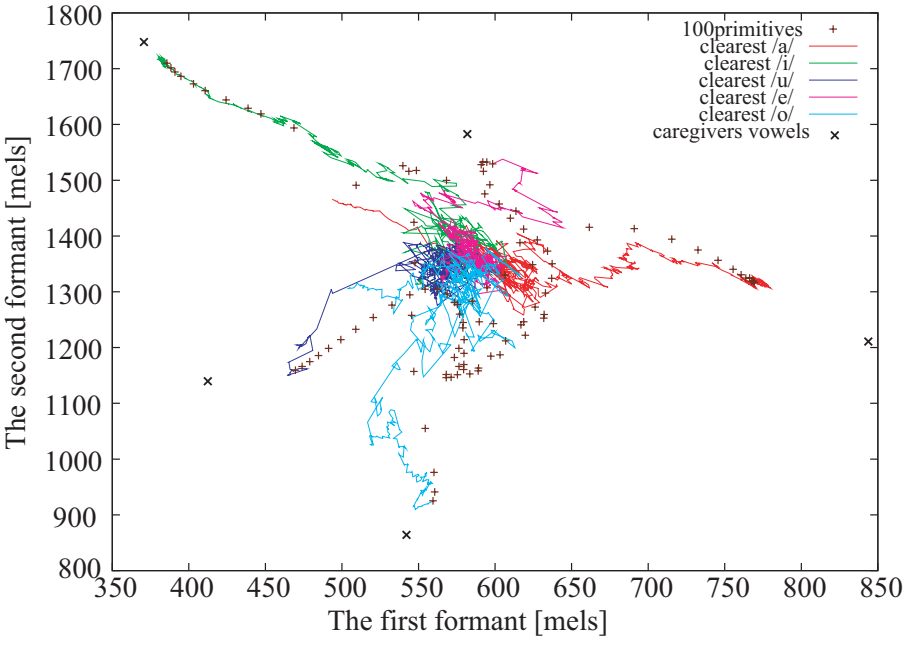
(a) beginning



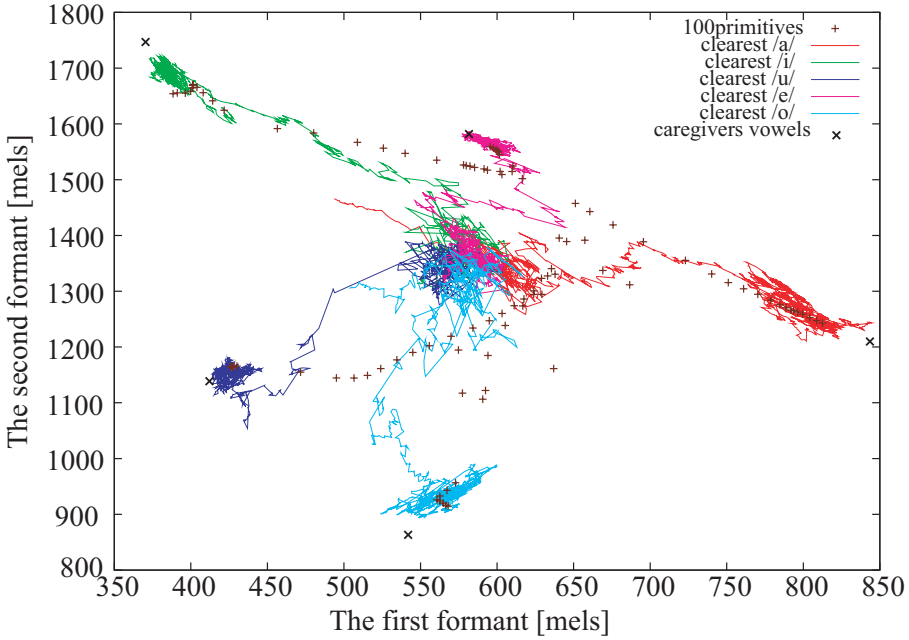
(b) 2000-th interaction



(c) 3000-th interaction



(d) 4000-th interaction



(e) 10000-th interaction

Figure 7: The trajectory of the center of Gaussian function $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_5$ and values of $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{100}$ under the condition with the imitation probability $p^I = 0.2$

primitive which enables the caregiver to imitate it again. Fig. 8 shows the transition of the success rate of the learner's imitation under $p^I = 0.2$. We can see that the learner became almost capable of imitating until the 4,000-th step in very high probability, namely about 0.9.

We calculated the averaged value of the selection probabilities of vowel primitives after 10,000 learning steps among ten trials under $p^I = 0.2$. Fig. 9 shows the summed values of $q_i(10,000)$ among primitives with the same p^S (vertical axis) in terms of p^S (horizontal one). It is, therefore, considered that the proposed method works well, making the learner acquire vowels easier for the caregiver to imitate in the case of the imitation probability $p^I = 0.2$.

Fig. 10 shows average performances at the 10,000-th step among ten trials under different p^I: discrepancy between the estimated centers of the imitation detectors and the corresponding caregiver's vowels (red curve) and the success rate of learner's imitation among ten trials (blue curve) while error bars indicate their standard deviations. We can see that the performances are well maintained even if p^I decreases to 0.1. This possibly shows the tolerance of the proposed method for variances in the characteristic of the caregiver whether he or she is more or less careful (imitative) of the learner.

5 Discussion concluding remarks

We have proposed a learning method that may explain the mechanism how infants can obtain their own vowels, exposed by their caregivers' mother tongues with less frequent (only 20%) imitative caregivers

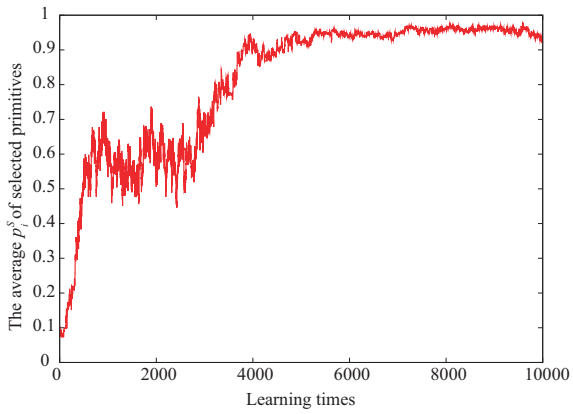


Figure 8: The trajectory of the success rate of the vowel imitation under the condition with the imitation probability $p^I = 0.2$

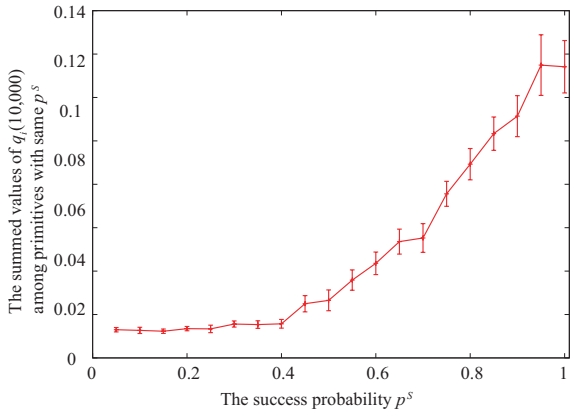


Figure 9: The summed values of $q_i(10,000)$ among primitives with the same p^S under the condition with the imitation probability $p^I = 0.2$

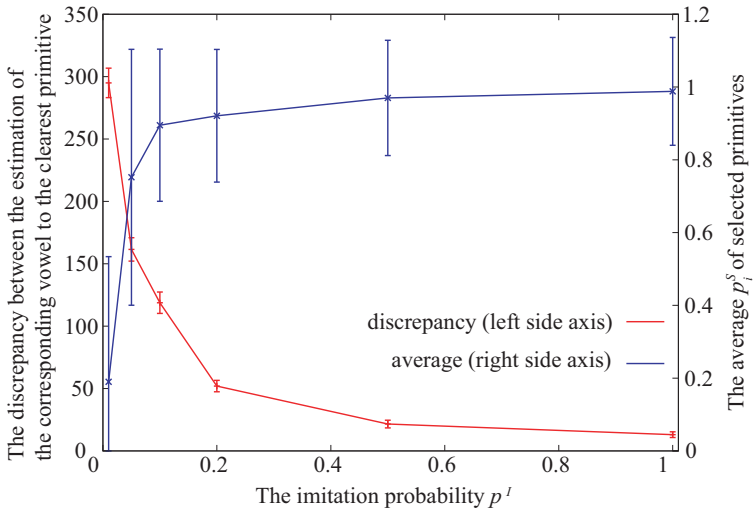


Figure 10: The average errors between the caregivers vowels and the center $\boldsymbol{\mu}_i$ of the Gaussian function $g_i(\boldsymbol{\mu}_i, \boldsymbol{S}_i)$ corresponding to each of the learner's vowels $/v_1/, \cdots, /v_5/$ and the success rate of the vowel imitation

as shown by Louis et al. [14]. As Kuhl et al. [15] have shown, the 6-month-old infants are capable to discern differences between the phonetic units of many languages, including languages they have never heard, but gradually their capabilities are tuned to their native languages. In order to model this process from a viewpoint of cognitive developmental robotics, we prepared many more vowel primitives on the learner side, and distributed their models around the caregiver's articulation region supposing that some among these models come to correspond to the caregiver's vowels while others do not during the learning process to simulate the infants developmental process. Although the proposed model seems to qualitatively explain the latter process of tuning in the suggestion of Kuhl et al. [15], we could improve the model to predict to what extent infants keep being capable to discern differences through development.

In addition, as shown in Fig. 8 and Fig. 9, the learner's utterances gradually shift to more vowel-like sounds and much clearer vowels (average $p_i^S = 0.9$) are uttered in the latter term of interaction. Such a change in the learner's utterances seems to resemble a certain aspect of infant development that has been reported in the literature [16] [17] where infant's utterances gradually shift from overlapped vowel clusters to separated (clearer) ones. In our model, the learner was able to select the utterances of the much clearer vowels which are actually easier for the caregiver to imitate because we assume that the learner selects the primitives according to degree of expectation e_i. It is uncertain whether or not infants really anticipate caregivers' imitations and reflect them in their utterances. However, some researches of developmental psychology report infants get interested in the caregiver's imitation and contingent response and prompt to them [18, 19, 20]. Therefore, we consider that further examinations on degree of expectation e_i become possible along with the advances in developmental psychology about the infant's behavior to the caregiver's imitation. Such investigations collaborated with developmental psychology are the important future issues to validate the plausibility of the proposed model.

Next, as shown in equation (8) and (9), σ is one of the parameters, which controls the anticipation of the caregiver's imitation. So, we analyze the relationship between the value of σ and the learning result. Fig. 11 shows the transitions of the discrepancy between the estimation centers of the Gaussian functions for the learner's clearest vowels and the formants of the caregiver's vowels under the condition with the imitation probability $p^I = 0.2$ and different σ. The red curve shows normal anticipation ($\sigma = 50$, the same result as the blue curve in Fig. 6), the blue curve shows weak anticipation ($\sigma = 200$), the green curve shows very weak anticipation ($\sigma = 500$), the pink curve shows strong anticipation ($\sigma = 1$). As shown in Fig. 11, the correspondence learning of the vowels does not go well in the earlier stage of learning when the anticipation is weak, and it also does not go well in the latter stage of learning when the anticipation is strong. This result indicates that the anticipation accelerates the correspondence learning, but too strong anticipation become to interfere it along with the learning progress. The curves and error bars in Fig. 12 show the average and the standard deviation of discrepancy of the clearest primitive in each category from its corresponding vowel at the end of the learning process with different parameters of p^I and σ. As shown in Fig. 12, level of the discrepancy with strong anticipation (pink curve) becomes less than those with weak anticipation (blue and green curves) and approaches one with

normal anticipation (red curve) as P^I is increased. These results indicate an important future issue of how the degree of anticipation can be adapted to the learning progress.

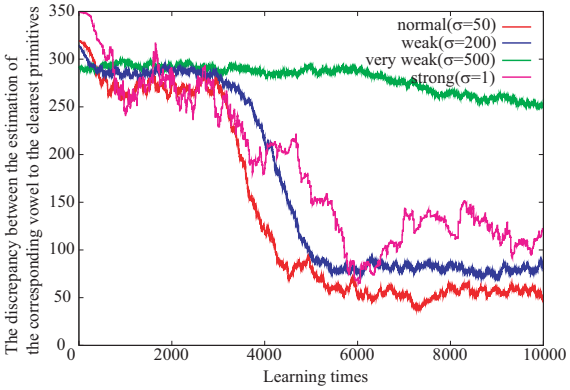


Figure 11: The transitions of the discrepancy between the estimation of the corresponding vowels to the clearest primitive of each category under the condition with the imitation probability $p^I = 0.2$ and different σ which is one of the parameters to calculate the the effect of the AMB $a_i(n)$

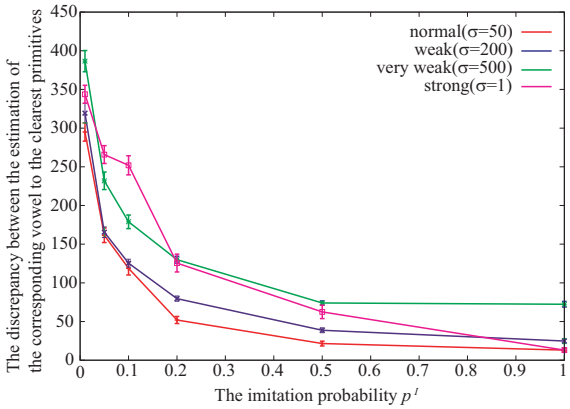


Figure 12: The average discrepancy between the estimation of the corresponding vowels to the clearest primitive of each category at the end of learning

The assumptions described in 2 seem very weak, maybe use: quite reasonable since they reflect the findings in developmental psychology. However, one big question is that we fixed 100 primitives of the learner's vowels, that means the capability of the articulation does not change while the change of the perception is modeled. That is, we ignore the emergence of more vowel-like utterances owing to the development of the vocal organ and change of utterable sounds for the infant in our study. This might be partially solved by representing the utterance by the learner that it expects to be imitated as not a discrete vector but a distribution on the formant space. Such distributions would be selected according to a degree of expectation q_i and utilized to explore better utterances. As a result, we can interpolate the utterance supposed to have been imitated. The extrapolation of the utterances can be involved in the process if we suppose that the learner's articulation skill develops. However, we need to know how this process happens in real infants.

Another issue is that we have not considered the effect of the AMB on the caregiver's side in the mapping process. In the Ishihara et al. [13] study, the caregiver's AMB and sensori-motor magnet affected the learning convergence of the infant robot. We may expect further acceleration of the learning, but there might be negative effects on the process. This and the development of the learner's articulation skill mentioned above are future work. Therefore, tuning of the hyper-parameters of learning such as ϵ in our model according to the history of the interaction would be necessary.

Final goal of our study is to realize interactions between a caregiver and an infant robot to verify the proposed method. The authors' group developed a research platform called CB2 [21] that has

different sensor modalities and body actions, however the articulation skill is insufficient. We need another research platform with a capability of richer articulation skills and realistic appearance so that the caregiver will feel like having a real interaction.

REFERENCES

[1] Giacomo Rizzolatti, Corrado Sinigaglia, and Frances Anderson, "Mirrors in the Brain - How Our Minds Share Actions and Emotions", Oxford University Press, 2008.

[2] B. de Boer, and W. Zuidema, "Multi-Agent Simulations of the Evolution of Combinatorial Phonology", Adaptive Behavior, vol. 18,No. 2, pp. 141-154, 2010.

[3] P.-Y. Oudeyer, "Phonemic Coding Might Result From Sensory-Motor Coupling Dynamics", Proceedings of the 7th international conference on simulation of adaptive behavior (SAB02), pp. 406-416, 2002.

[4] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno, "Continuous Vocal Imitation with Self-organized Vowel Spaces in Recurrent Neural Network", Proceedings of IEEE International Conference on Robotics and Automation, pp. 4438-4443, Kobe, May, 2009

[5] Kotaro Fukui, Kazufumi Nishikawa, Shunsuke Ikeo, Masaaki Honda, and Atsuo Takanishi, "Development of Human-like Sensory Feedback Mechanism for an Anthropomorphic Talking Robot", IEEE International Conference on Robotics and Automation 2006, pp. 101-106, 2006.

[6] P. K. Kuhl, and B. T. Conboy, and S. Coffey-Corina, and D. Padden, and M. Rivera-Gaxiola, and T. Nelson, "Phonetic learning as a pathway to language: new data and native language magnet theory expanded(NLM-e)" Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 363, No. 1493.(12 March 2008), pp. 979-1000, 2008.

[7] N. Masataka and K. Bloom, "Acoustic Properties That Determine Adult's Preference for 3-Month-Old Infant Vocalization." Infant Behavior and Development, vol. 17, pp. 461-464, 1994.

[8] M. Peláez-Nogueras, J. L. Gewirtz, and M. M. Markham, "Infant vocalizations are conditioned both by maternal imitation and motherese speech." Infant behavior and development, vol. 19, pp. 670, 1996.

[9] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida."Cognitive developmental robotics: a survey", IEEE Transactions on Autonomous Mental Development, vol. 1, no. 1, pp. 12-34, 2009.

[10] Yuichiro Yoshikawa and Minoru Asada and Koh Hosoda and Junpei Koga, "A Constructivist approach to infants' vowel acquisition through mother-infant interaction." Connection Science, vol. 15, no. 4, pp. 245-258, December 2003.

[11] Patricia K. Kuhl, "Plasticity of development", chapter 5 Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech., pp. 73-106, MIT Press, 1991.

[12] Katsushi Miura, Minoru Asada and Yuichiro Yoshikawa, "Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories." RSJ Advanced Robotics Special Issue on Imitative Robots, vol. 21, no. 13, pp. 1583-1600, September 2007.

[13] Hisashi Ishihara, Yuichiro Yoshikawa, and Minoru Asada. "Caregiver's Auto-mirroring and Infant's Articulatory Development Enable Vowel Sharing" Proceedings of Ninth International Conference on Epigenetic Robotics, pp. 65-72, 2009.

[14] Julie Gros-Louis, Meredith J. West, Michael H. Goldstein, and Andrew P. King. "Mothers provide differential feedback to infants' prelinguistic sounds" International Journal of Behavioural Development, vol. 30, No. 6, pp. 509-516, 2006.

[15] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom. "Linguistic experience alters phonetic perception in infants by 6 months of age" Science, vol. 255, pp. 606-608, 1992.

[16] K. Ishizuka, R. Mugitani, H. Kato, and S. Amano. "Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infant." Journal of Acoustic Society of America, vol. 121, no. 4, pp. 2272-2282, 2007.

[17] P. K. Kuhl, and A. N. Meltzoff. "Infant vocalizations in response to speech: Vocal imitation and developmental change." Journal of Acoustic Society of America, vol. 100, pp. 2415-2438, 1996.

[18] Watson J. S. "Contingency perception in early social development." In T. M. Field and N. A. Fox, editors, Social Perception in Infants, pp. 157-176, N. J.: Ablex, Norwood, 1985.

[19] Kathleen Bloom, Ann Russella1, and Karen Wassenberg. "Turn taking affects the quality of infant vocalizations." Journal of Child Language, vol. 14, pp. 211-217, 1987.

[20] M. Peláez-Nogueras, J. L. Gewirtz, and M. M. Markham. "Infant vocalizations are confitioned both by maternal imitation and motherese speech." Infant behavior and development, vol. 19, pp. 670, 1996.

[21] T. Minato, Y. Yoshikawa, T. Noda, S. Ikemoto, H. Ishiguro, and M. Asada. "CB^2: A Child Robot with Biomimetic Body for Cognitive Developmental Robotics" Proc. of the IEEE/RSJ International Conference on Humanoid Robots, CD-ROM, 2007.