# Realizing being imitated: vowel mapping with clearer articulation

Katsushi Miura[1)2)], Yuichiro Yoshikawa[1)], and Minoru Asada[1)2)]
[1)] *JST ERATO Asada Synergistic Intelligence Project (www.jeap.org)*
[2)] *Graduate School of Engineering, Osaka University*
{*miura,yoshikawa,asada*}*@jeap.org*

*Abstract*— **The previous approach to vowel imitation learning between a caregiver and an infant (robot) [1] has assumed that the robot can segment the caregiver's utterance into its phoneme category, where the caregiver always imitates the robot utterance. However, in real situations, the caregiver does not always imitate the robot utterance, nor the robot does have the phoneme category (no segmentation capability). This paper presents a method to solve these issues, a weakly-supervised learning along with auto-regulation, that is active selection of action and data with underdeveloped classifier. To cope with not-always imitation problem, a weakly-supervised learning method is applied that is capable to handle incompletely segmented samples (not perfectly imitated voices). Further, the regulation classifier of the imitated voices is recursively applied in order to select good vocal primitives and to segment caregiver's imitated voices that improve the performance of the classifier itself. The simulation results are shown and the future issues are given.**

*Index Terms*— **Interaction, Imitated voice, Weakly-supervise, Self-supervise**

## I. Introduction

Humanoid robots have been expected to communicate with humans through the modals such as voice and gestures humans utilize. However, due to the difference in physical structure between them, it is difficult for humanoid robots to understand and reproduce the observed humans' actions as they are. On the other hand, human infants seem to successfully solve the same issue with the different articulation structure in the developmental process of language faculty.

Previous studies have demonstrated that a population of computer-simulated agents with a vocal tract and cochlea could self-organize shared vowels through imitating each other [2], [3]. Fukui et al. [4] showed that the robot can acquire consonant sounds by reducing the error between the pressure of human voice and the robot consonant production. However, these studies ([2], [3], [4]) 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. Developmental psychologists show that infant's vowel-like cooing tends to invoke utterances by its mother's imitation (Masataka and Bloom [5]) and that maternal imitation effectively reinforces infant vocalization (Pel'aez et al. [6]). Therefore, we may assume that the interactions between mother and infant enables the infants to understand and imitate the mother's voice, and our group has applied this idea to designing the vocal robots and interaction with their caregivers in the context of cognitive developmental robotics [7].

Yoshikawa et al. [8] constructively showed that imitation by the caregiver in response to infant's vocalization with vowels plays an important role in vowel acquisition. Considering well-known "perceptual magnet effect" [9] by which person's perception of phonemes is biased to his/her own category, Miura et al. [1] showed that the utterance of baby robot could be leaded to clear vowel sound through the mutual imitation with its caregiver. In these studies, it was assumed that the caregiver always imitated single vowel-like sound of the robot but did not utter anything else. In other words, the robot did not need to find the sound to imitate the robot's demonstrated voice in the caregiver's responding voice. This can be against the real mother-infant interaction where the mother often adds or modifies their voices which do not always correspond to the precedent infant's voices exactly. What's the worse is that the mother might sometimes fail to listen to and imitate a infant's voice since the infant's vocalization skill is not matured yet.

In this paper, we consider a more realistic scenario of infant - caregiver interaction in the context of joint attention with a number of objects on the table between them. In such a scene, the caregiver guesses a word from the incomplete utterance of the infant, and may confirm if the guessed word is correct or not. In such interactions, the percentage of the imitation of all or part of the robot's syllables by the caregiver side is still high even though she sometimes fails to imitate due to the ambiguity of the robot's utterance. Thus, the caregiver does not always imitate the infant utterance, that releases the assumption in the previous studies.

To cope with the not-always imitation issue, we propose a method a weakly-supervised learning along with auto-regulation, that is active selection of action and data with underdeveloped classifier. A weakly-supervised learning method is applied which is capable to handle incompletely segmented samples (not perfectly sounds to imitate the robot's demonstrated voice). Further, the regulation classifier of imitated voices is recursively applied in order to select better vocal primitives and to segment caregiver's responding voices that improve the performance of the classifier itself. In the rest of the paper, we first explain detailed setting of the vocal interaction. Then, we introduce the proposed method of a weakly-supervised learning along with auto-regulation, that is active selection of action and data with underdeveloped classifier. Finally, we show the experimental results and future issues.

## II. How the robot and a caregiver interact?

Suppose that a learner (robot) and a caregiver (human) repeat to alternately utter a sequence of a number of phonemes. We assume that there are $N_h$ vowels in the caregiver's language system. Let $/v_i/, (i = 1, \cdots, N_h)$ denote the $i$-th vowel of the language. On the other hand, the robot is assumed to have $N_h$ groups of phonemic primitive to utter vowels. Each of these groups consists of $N_r^{/v_i/}$ primitives which can be recognized as $/v_i/$ in different probabilities. The $j$-th primitive for $/v_i/$ is denoted as $/v_{i,j}/$ $(j = 1, \cdots, N_r^{/v_i/})$. The probability of being successfully recognized as $/v_{i,j}/$ is called propagation rate and denoted by $P^{/v_{i,j}/}$. It is assumed that the robot does not know the correspondence between caregiver's $/v_i/$ and robot's $/v_{i,j}/$ nor the probabilities $P^{/v_{i,j}/}$.

In every interaction, the robot chooses the $N_a$ primitives from its all primitives and utters them. The $k$-th primitive chosen by the robot at the $t$-th interaction is labeled as $/R_k^t/, (k = 1, \cdots, N_a)$. How to choose primitives is explained in section Ⅲ. The caregiver imitates a part or all of the robot's primitives. Here, we assume the caregiver intends to imitate the robot's $k$-th vowel according to the imitation rate $P(0 \leq P \leq 1)$. The $k$-th vowel uttered by the caregiver by the $t$-th interaction is labeled as $/H_k^t/$. In other words, the caregiver chooses corresponding vowels to the robot's choice at the probability of $PP^{/R_k^t/}$ or another at the probability of $(1-PP^{/R_k^t/})/(N_h-1)$ for each vowel, and utters these $N_a$ vowels in order.

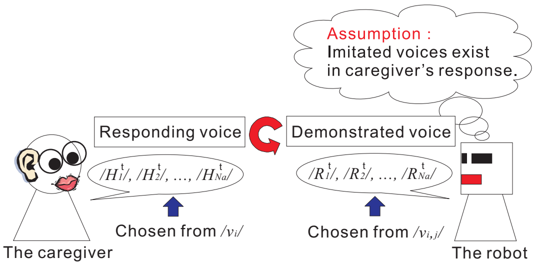


Fig. 1. An overview of the interaction

The robot listens to the caregiver's responding voice and obtains a sequence of the continuous sound pressure, $S^t = \{s(1), \cdots, s(d^tF)\}$, where $s(l)$ is the sound pressure when $\frac{l}{F}$ seconds has passed from the beginning, $d^t$ is the duration of the caregiver's utterances in the $t$-th interaction and $F$ is the sampling frequency. Through T steps interaction, the robot obtains T sets of its own labels of demonstrated voice and the response sound by the caregiver, that is $\{R_1^t, \cdots, R_{N_a}^t, S^t\}$, $(t = 1, \cdots, T)$. It uses them to learn $N_r^{/v_i/}$ detectors each of which identifies the caregiver's responding sound to imitate $/v_{i,j}/$ with weakly-supervised learning.

## III. The method

Through the interaction with the caregiver, the robot acquires imitation detectors from the caregiver's responding voice. An imitation detector $f^{/v_{i,j}/}$ is defined for each vowel primitive of the robot $/v_{i,j}/$ and used for segmenting the caregiver's responding voice into periods including corresponding sound to $/v_{i,j}/$ and the other ones. The acquired imitation detector is used not only for evaluating whether the robot's demonstrated voices are imitated but also biasing learning data to involve successful experiences. Consequently, learning of classifier is auto-regulated. An incremental learning system is proposed based on the weakly-supervised learning

### A. weakly-supervised learning along with auto-regulation

An imitation detector of $/v_{i,j}/$ can be described by a classifier $f^{/v_{i,j}/}$ which outputs the likelihood of the label $/v_{i,j}/$ for the current sound feature $\boldsymbol{m}^t$ that can be extracted from the caregiver's responding voice $S^t$. The basic idea to let the robot acquire the imitation detectors is to let it have a brief that its demonstrated voices are always imitated by the caregiver. We suppose that it is partially true that a mother tends to imitate her infant's behavior not only in the case where they enjoy imitating each other but also in the case where they try to share their attention with verbal cues, such as the name of a topic. In this case, we apply weakly-supervised learning since it presumes that the input data $S^t$ can be labeled by its all precedent primitives $/R_1^t/, \cdots$, or $/R_{N_a}^t/$. As a result locates where the corresponding part is. Though, the input data do not sometimes involve the corresponding parts to the labels due to the assumption that the caregiver does not always imitate them.

Fig. 2 shows an overview of updating system of imitation detectors with weakly-supervised learning. The robot calculate the sound feature $\boldsymbol{m}_n^t$ that can be extracted from the caregiver's responding voice $S^t$. Imitation detectors evaluate whether the sound feature $\boldsymbol{m}_n^t$ corresponds to the one of the robot's demonstrated voice $/R_k^t/$. The robot repeats the interaction $T$ times and updates the imitation detector for each primitives from $T$ sets of training data consisting of pairs of sound feature $\boldsymbol{m}_n^t$ and the robot's vocal primitive $/R_k^t/$.

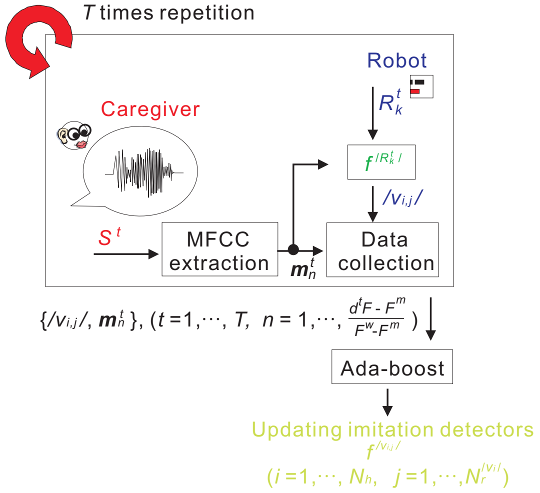

T times repetition
Robot
$R_k^t$
Caregiver
$f^{/R_k^t/}$
$/v_{i,j}/$
$S^t$
MFCC extraction
$\boldsymbol{m}_n^t$
Data collection
$\{/v_{i,j}/, \boldsymbol{m}_n^t\}, (t = 1, \cdots, T,\ n = 1, \cdots, \frac{d^tF - F^m}{F^w - F^m})$
Ada-boost
Updating imitation detectors
$f^{/v_{i,j}/}$
$(i = 1, \cdots, N_h,\ j = 1, \cdots, N_r^{/v_i/})$

Fig. 2. An overview of weakly-supervised learning

*1) MFCC (Mel-Frequency Cepstral Coefficient):* We used the well-known MFCC as the sound feature of the caregiver's utterances. The robot extracts the characteristic quantity of 25 dimensions ($\boldsymbol{m}_n^t = [m_n^t(1), \cdots, m_n^t(25)]^T, (n = 1, \cdots, \frac{d^tF - F^m}{F^w - F^m})$). This extraction of MFCC is proceeded with a sampling window $F^w$ and a moving window $F^m$ for the caregiver's utterance $S^t$ in $d^t$ seconds. The sound features that we used in this research consist of 25 dimensional vectors of which the first element represents the power of the sound and from the second to the 13th elements represent powers of the spectrum on to the Mel scale, using triangle overlapped windows and from 14th to 25th represents the gradient of powers of the spectrum.

*2) imitation detector:* By the imitation detectors, the robot segments the caregiver's responding voice and judge vowel labels every $\frac{F^w}{F}$ seconds. The primitive $/v_{i,j}/$, that corresponds to MFCC 25 dimension value $\boldsymbol{m}_n^t$ (extracted from $t$ turns interaction), was chosen from own vowel labels $/R_k^t/$ by the imitation detector $f_{/R_k^t/}(\boldsymbol{m}_n^t)$.

$$/v_{i,j}/ = \underset{/R_k^t/}{\arg\max} f^{/R_k^t/}(\boldsymbol{m}^t) \tag{1}$$

When all the values of $f_{/R_k^t/}(\boldsymbol{m}_n^t)$ are less than the threshold value (=0), $\boldsymbol{m}_n^t$ will be recognized as a noise belonging to no vowel sound label. The robot acquires $T$ sets of training data consisting of input as $\boldsymbol{m}_n^t, (t = 1, \cdots, T)$ and output as $/v_{i,j}/$ from $T$ times interactions, and use them for updating the imitation detector. Then, it updates the imitation detector $f_{/R_k^t/}(\boldsymbol{m}_n^t)$ by Ada-boost (see below). In addition, the robot improves performance of imitation detectors by repeating weakly-supervised learning $K$ times.

*3) Ada-boost:* Ada-boost is a method to learn a classifier $f(x)$ so that it shows the high performance of distinguishing the voice imitating it by combining several weak learners $f_n(x), (n = 1, \cdots, N)$ which has low-ability and return binary value encoding true or false. $N$ is the number of weak learners here. In weakly-supervised learning, we assume $\boldsymbol{m}_n^t$ as input signal, and $/v_{i,j}/$ which is judged by the imitation detectors as output. We suppose the robot acquires higher-ability imitation detector $f^{/v_{i,j}/}$, which detects imitated sound from the caregiver's responding voice in every primitive $/v_{i,j}/$ of the $N_r$ units.

*4) design of weak learners for filtering:* As implemented in a study by Viola and Jones [10] about a face recognition method by using Ada-boost, rectangular features are used. It is calculated from rectangular areas consisting of contiguous pixels from the various positions in the image picture. A classifier to distinguish whether a face is included in an image was acquired by combining several weak learners, which calculate simple addition and subtraction of pixel values. There is a limitation in the selection of rectangles. Only the combination of two, three, or four same size's and contiguous rectangles is used.

We constructed original weak learners referred to the method of Viola and Jones in this study, and used MFCC of caregiver's utterance ($\boldsymbol{m}_n^t$ with 25 dimensions) as an image of the 25×1 pixels. We designed weak learners, which have a filtering ability, by combining in rectangular areas of four patterns (as $m_n^t(k), (k = 1, \cdots, 25)$ pixels). The rectangular areas combination of these four patterns are shown in Fig. 3. (a) has only one rectangle, and becomes $m_n^t(l)$,$(l = 1, 2, \cdots, 25)$. (b) is a difference of pixel values from two rectangles that do not contact each other.

$$m_n^t(l) - m_n^t(l') \tag{2}$$

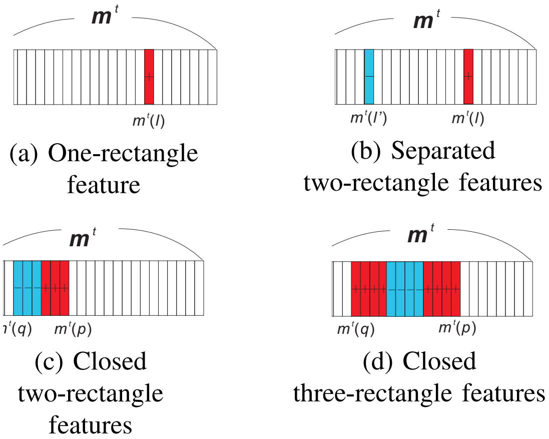
Here, it is $l > l' + 1$, $l = 3, \cdots, 25$, $l' = 1, \cdots, 23$. (c) is a difference of two contiguous rectangle's pixel value that has an equal area.

$$\sum_{l=p}^{p+\frac{q-p-1}{2}} m_n^t(l) - \sum_{l=p+\frac{q-p+1}{2}}^{q} m_n^t(l) \tag{3}$$

Here, it is $p > 1, q < 25, q - p + 1 = 2k + 1 > 0, k = 1, 2, \cdots, 11$. (d) is a difference of three equal rectangle's pixels value, which have contacts with each other.

$$\sum_{l=p}^{\frac{q-p-2}{3}} m_n^t(l) - \sum_{l=\frac{q-p+1}{3}}^{p+\frac{2q-2p-1}{3}} m_n^t(l) + \sum_{l=p+\frac{2q-2p+2}{3}}^{q} m_n^t(l) \tag{4}$$

Here, it is $p > 1, q < 25, q-p+1 = 3k > 0, k = 1, 2, \cdots, 7$. The number of weak learners is 534 in total - with each combination of (a) 25 ways, (b) 253 ways, (c) 156 ways, and (d) 100 ways. Now, we will search for the most adequate combination from the 534 ways by Ada-boost.



(a) One-rectangle feature

(b) Separated two-rectangle features

(c) Closed two-rectangle features

(d) Closed three-rectangle features

Fig. 3. Example of rectangle features

*5) active action and data selection:* The robot can count with the imitation detectors, whether imitations are included in caregiver's reply, i.e. when the caregiver's voice to imitate $/v_{i,j}/$ was uttered in the interaction. The utterance probability $Q_K^{/v_{i,j}/}(0 \leq Q_K^{/v_{i,j}/} \leq 1)$ of every primitive $/v_{i,j}/$ in $K$ time's supervised learning is defined in terms of the number of imitation times, when $/v_{i,j}/$ was uttered in $K$ time's weakly-supervised learning (if $K = 1$, $Q_1^{/v_{i,j}/}$ is same value in every primitive $/v_{i,j}/$). In this study, the number of imitation times is categorized by each vowel $/v_i/$. When the robot utters vowel sound $/v_i/$ in $K$ time's self-supervised learning, the probability of utterance $Q^{/v_{i,j}/}$ is chosen according to the following formula.

$$Q^{/v_{i,j}/} = \frac{I^{/v_{i,j}/}}{\sum_{i,j} I^{/v_{i,j}/}} \tag{5}$$

Here, $I^{/v_{i,j}/}$ is the number of imitation times for an utterance $/v_{i,j}/$ in $(K-1)$ time's supervised learning. We can expect that the number of imitation times increase by higher propagation rate $Q^{/v_{i,j}/}$. This means that the number of robot's utterances will increase by repeating self-supervised learning, when the primitive utterance has a high transmission rate $Q^{/v_{i,j}/}$.

*6) initial learning:* Since the initial value of each imitation detector is unknown for the robot, the robot must acquire the initial value of every imitation detector without self classification system in initial weakly-supervised learning. Fig 4 shows an overview of acquiring initial values of imitation detectors with initial weakly-supervised learning. In this initial weakly-supervised learning, since the robot cannot segment the caregiver's responding voice, the robot calculates the average value with 25 dimensions $\boldsymbol{m}^t$ of the whole $S$, and used it as input of the learning after extraction of $\boldsymbol{m}_n^t$.

$$\boldsymbol{m}^t = \frac{\sum_n \boldsymbol{m}_n^t}{\frac{d^t \cdot F - F^m}{F^w - F^m}} \tag{6}$$

From $T$ times interactions, the robot obtains the $T$ sets of training data consisting of pairs of sound feature $\boldsymbol{m}^t$ and the robot's vocal primitive $/R_1^t/, \cdots$, or $/R_{N_a}^t/$.

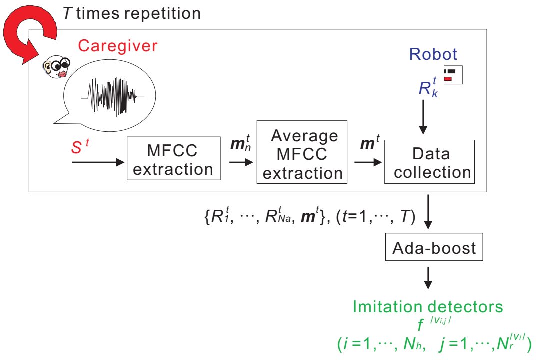


Fig. 4. An overview of initial weakly-supervised learning

## IV. EXPERIMENTS

There are two learning tasks for the robot in this interaction. The first task is to find the imitation sound of the caregiver and to acquire the correspondences between own primitive vowel sounds and caregiver's vowels. The second task is to increase the utterance probability of the robot's primitive with high propagation rate $P^{/v_{i,j}/}$.

### *A. setting*

On computer simulation, we implemented an experiment in which vowel label $/R_k^t/$ uttered by the robot and vowel label $/H_k^t/$ returned by the caregiver were chosen repeatedly. In this simulation, we chose randomly the one of caregiver's utterances which were recorded before the experiment and decided the robot's vowel label $/R_k^t/$ corresponding to the caregever's one by using the imitation rate $P$ and propagation rate $P^{/v_{i,j}/}$. The number of uttered syllables $N_a$ is three in this interaction. The caregiver is supposed to be Japanese who has Japanese five vowels ($N_h = 5$, $/v_i/ = /a/, /i/, /u/, /e/$, or $/o/$) as vowel primitives, and the robot has three primitives for each Japanese vowel and therefore totally 15 primitives. The propagation rates are set as $P^{/v_{i,1}/} = 0.4$, $P^{/v_{i,2}/} = 0.7$, and $P^{/v_{i,3}/} = 1.0, (i = 1, \cdots, N_h)$.

A Japanese male was asked to utter series of syllables consisting of three vowels only (hereafter V-syllables) or three vowels with consonants (hereafter CV-syllables) that were used as the responding sound from a caregiver. His all utterances were recorded in sampling frequency 16000[Hz]. The types of recorded sounds and the number of them are listed in Table I. The numbers of the prepared set of V-syllables and CV-syllables were 1000. Since the number of the possible order of three Japanese vowels is 125, he uttered each V-syllables eight times. To order of CV-syllables were randomly chosen from possible candidates. He also asked to utter five single V-syllables 10 times and 39 single CV-syllables 10 times which were used for evaluation.

TABLE I
VERSIONS OF CAREGIVER'S UTTERANCES

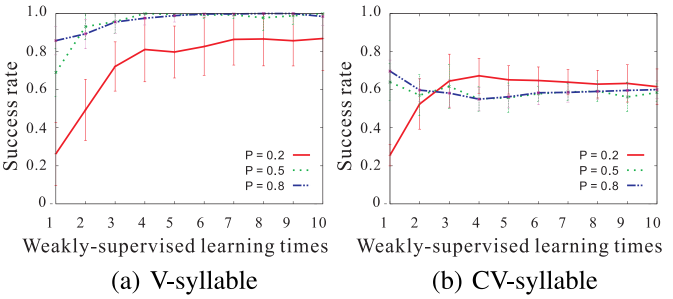
| | |
|---|---|
| Three V-syllables | 1000 |
| Three CV-syllables | 1000 |
| A single V-syllable for estimation of learning machines | 50 |
| A single CV-syllable for estimation of learning machines | 390 |

The weakly-supervised learning was repeated 9 times after initial one proceeded. In the initial weakly-supervised learning, the number of learning times $T$ was 1000 while it was 100 in the latter ones. Then, we evaluated the ability of the acquired classifier for imitated sounds. The acquired classifiers were applied, every time window of $\frac{F^w}{F}$ seconds, for the prepared data set $S$ of single V-syllables and CV-syllables. A vowel that was most frequently identified by using acquired classifiers through time windows was chosen for each $S$. We then calculate the success rate of this classification whether the chosen vowel corresponds to the true label existing in $S$. Note that the stable part of the CV-syllable was mainly evaluated in this analysis since a Japanese CV-syllable consists of a transition part, i.e., consonant, and a longer stable part, i.e., a vowel. In all experiments, the sampling window $F^w$ is 512 data points (about 0.03 seconds), and the moving window $F^m$ is 256 data points.

### B. experiment result

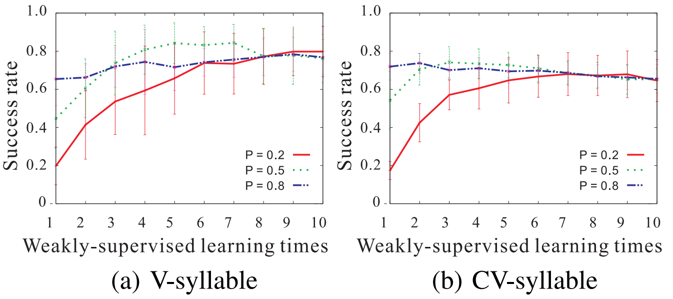
We ran 10 trials of learning simulation under various condition where the imitation rate $P$ = 0.2, 0.5, and 0.8. In each trial, the robot iterated nine weakly supervised learning processes after initial ones. We calculated the averages and standard deviations of success rates for both cases where V-syllables (see Figs. 5) and CV-syllables (see Figs.6) were used for the training data. Fig. 5 (a) and Fig. 6 (a) shows the success rates on the validation data consisting of single V-syllables while Fig. 5 (b) and Fig. 6 (b) shows those consisting of single CV-syllables. Note that in these figures, red, green, and blue curves indicates the result under different imitation rate $P$ = 0.2, 0.5, and 0.8 respectively.

Figs. 5 and Figs. 6 show that if imitation rate $P$ is 0.8, the robot succeeded in learning almost correct imitation detectors only by initial weakly-supervised learning step (see blue curves in Figs. 5 and Figs. 6). On the other hand, the ability of the imitation detectors obtained only by the initial step is low (see left edge of red curves in Figs. 5 and Figs. 6), if imitation rate $P$ is 0.2. However, through repeating weakly-supervised learning, it could acquire imitation detectors with high success rates, even if the imitation rate $P$ is low (see red curves in Figs. 5 and Figs. 6).



(a) V-syllable (b) CV-syllable

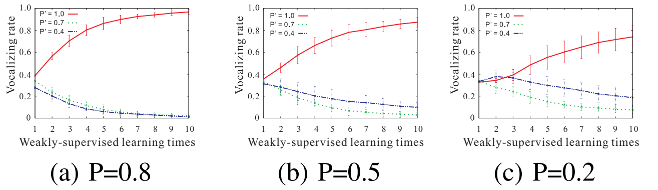
Fig. 5. Comparison of success rate of detecting being imitated when the caregiver replied with V-syllables and the propagation rate $P^{/v_{i,j}/} = 1.0$



(a) V-syllable (b) CV-syllable

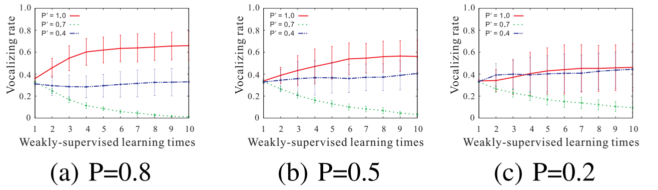
Fig. 6. Comparison of success rate of detecting being imitated when the caregiver replied with CV-syllables and the propagation rate $P^{/v_{i,j}/} = 1.0$

We then analyzed how the robot's utterances were changed according to the mechanism of active selections through the interaction. Figs. 7 show the average changes in utterance frequency of each primitive when V-syllables were used as the caregiver's responding sound. Figs. 7 show the transitions of selected frequencies along with learning steps for each group of syllables with the same propagation rate under different imitation rates (a) $P$ = 0.8, (b) $P$ = 0.5, (c) $P$ = 0.2. On the other hand, Figs. 8 show those when CV-syllables were used as the caregiver's responding sound.



(a) P=0.8 (b) P=0.5 (c) P=0.2

Fig. 7. The average changes in utterance frequency of each primitive when when the caregiver replied with V-syllables



(a) P=0.8 (b) P=0.5 (c) P=0.2

Fig. 8. The average changes in utterance frequency of each primitive when when the caregiver replied with CV-syllables

From the results of Figs. 7 and Figs. 8, we found that the frequency of primitives with the highest propagation rates increased along with the repetition of weakly-supervised learning. The primitives with the highest propagation rates are regarded as the corresponding vowels. Considering the results that the success rates of classification could increase (see Figs. 5 and Figs. 6) and that the corresponding primitives become to be more selected (see Figs. 7 and Figs. 8). Learning of vowel correspondences and active selection of the primitives are mutually encouraged through the interaction with the caregiver. In other words, it could become realizing being imitated by the proposed method.

## V. Discussion and conclusion

In the experiment, through the interaction with a caregiver who responds to the robot's demonstrated voice, the robot acquired imitation detectors which could segment the caregiver's responding voice into each vowel. The success rates of imitation detectors acquired through the interaction via V-syllables was higher than those acquired via CV-syllables (compare Fig. 5 (a) and Fig. 6 (a)). This was regarded to be caused by the fact that the transition part of CV-syllables, that is consonants, obstructed learning of the correspondences. Also, frequency of utterance with the highest propagation rate become higher when the caregiver responds to it with V-syllables than when he did with CV-syllables. Applying these results for the mother-infant interaction, it is considered that the infant could acquire the corresponding vowels to mother's ones more easily, if the caregiver utters vowels or stable part of vowels with consonants as long as possible and the infant more frequently utters primitives which are easier for him to imitate. In other words, the experimental result might imply that mother's response with slower utterance, like a motherese, is effective to the infant's vowel acquisition process in the meaning that it would help infant's learning of correspondence.

As shown in the experimental results in Fig. 7 (c) and Fig. 8 (c), the robot succeeded in acquiring the correspondences with lower imitation rate such as $P = 0.2$, which is considered as the chance level. The reason why might be that the probability of hearing corresponding syllables to an utterance in the caregiver's responding sound is actually higher than the imitation rate $P$. This could happen because the occasional correspondence might be happened when the caregiver did not imitate the other part's in the series of the robot's utterance. In the mother-infant interaction, mother could utter with more than a few syllables. Therefore, the probability that imitation sound are included in a mother's response could be higher than chance level. By utilizing this occasional imitation learning, the infant can more easily acquire the vowel correspondences to his/her mother, and increase vowel-like utterances in the interaction.

Furthermore through repeating weakly-supervised learning, the robot could acquire imitation detectors which have higher classifying ability and become choosing most natural primitives frequently, even though a imitation rate $P$ is low, which might indicate the effectivity of the proposed method. In such a way, the current simulation demonstrates that learning of vowel correspondences and selecting of primitives with highest propagation rates seemed to be mutually encouraged, in other words, the robots could become realizing being imitated. However, we have not considered the change of the caregiver's responding voice along with the change of the robot's demonstrated voice since the caregiver's way of responding was fixed. To consider such changes of the caregiver, we must examine the real interaction with a human caregiver by using a real vocal robot.

## References

[1] hKatsushi 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.

[2] B. de Boer, "Self-organization in vowel systems", Journal of Phonetics, vol. 28, pp. 441-465, 2000.

[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] 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.

[5] 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.

[6] 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.

[7] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro and Yasuo Kuniyoshi, "Cognitive Developmental Robotics As a New Paradigm for the Design of Humanoid Robots.", Robotics and Autonomous System, vol. 37, pp. 185-193, 2001.

[8] 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.

[9] 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.

[10] Paul Viola and Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features." IEEE Computer Vision and Pattern Recognition, 2001.