A Constructive Model of Mother-Infant Interaction towards Infant's Vowel Articulation

Yuichiro Yoshikawa
Osaka University
Yamada-Oka 2-1, Suita
Osaka 565-0821 Japan
yoshikawa@er.ams.eng.osaka-u.ac.jp

Junpei Koga
Osaka University
Yamada-Oka 2-1, Suita
Osaka 565-0821 Japan
koga@er.ams.eng.osaka-u.ac.jp

Minoru Asada
HANDAI Frontier Research Center,
Osaka University
Yamada-Oka 2-1, Suita
Osaka 565-0821 Japan
asada@ams.eng.osaka-u.ac.jp

Koh Hosoda
HANDAI Frontier Research Center,
Osaka University
Yamada-Oka 2-1, Suita
Osaka 565-0821 Japan
hosoda@ams.eng.osaka-u.ac.jp

Abstract

Human infants seem to develop to acquire common phonemes to adults without the capability to articulate or any explicit knowledge. To understand such unrevealed human cognitive development, building a robot which reproduces such a developmental process seems effective. It will also contribute to a design principle for a robot that can communicate with human beings. This paper hypothesizes that the caregiver's parrotry to the coo of the robot plays an important role in the phoneme acquisition process based on the implication from behavioral studies, and propose a constructive model for it. We validate the proposed model by examining whether a real robot can acquire Japanese vowels through interactions with its caregiver.

1. Introduction

In order for a new born agent to communicate by vocalization with senior members of a society, it should acquire common phonemes to the seniors as primitives. A human infant seems to acquire phonemes of its mother tongue and finally whole mother tongue without explicit knowledge about the relationship between its sensorimotor systems and phonemes, even though it has not developed to be able to articulate adult phonemes as they are. Although there are some studies on the infant behavior and development, details of its process have not been revealed yet. The aim of this study is to suggest a constructive model of the infant's cognitive developmental process by building a real robot that develops to acquire human phonemes.

Since an infant does not seem to have any explicit knowledge about the relations between its sensorimotor system and phonemes, it needs to learn the relation through interactions with its environment, namely its caregiver who teaches phonemes. Therefore, we have two main design issues to learn phonemes; 1) what kind of mechanism for interaction should be embedded in the robot and 2) what kind of reaction should be taken by the caregiver.

There are a number of studies on infant development of vocalization learning which give us some suggestions. Until about the second month, an infant begins cooing both in speech-like and unspeech-like manners. It is reported that maternal imitation of infant's cooing, that is parrotry vocalization, increases vocalization rates of a 3-month-infant (Peláez-Nogueras et al., 1996) and its speech-like cooing tends to lead utterances of its mother (Masataka and Bloom, 1994). Based on these observations, we conjecture that the caregiver's vocalization responded to the infant's cooing reinforces infant's articulation of cooing which can be interpreted as phonemes, and that the caregiver's parrotries give instructions about the correspondence between cooing and the caregiver's phonemes and about the acoustic information about phonemes. In order to verify these conjectures, we propose a preliminary, constructive model of mother-infant interaction by the caregiver-robot interaction based on embedding a random vocalization mechanism in it and letting the caregiver parrot its vocalization, which works so that the robot can succeed in the phoneme acquisition.

Our experimental robot has the similar structure to articulate with the one used in the previous robotics study (Higashimoto and Sawada, 2002), which articulates by deforming a silicon vocal tract. Because of the limitation in such a structure, we cope with only vowels as the first step. Two innate mechanisms are embedded in it. One is an extractor of formants which are well known effective sound features to distinguish vowels. Another is a learning mechanism which consists of auditory and articulation layers and connections between them. The auditory layer is for clustering formants of the caregiver by self-organization (Kohonen, 1984) while the articulation one is for clustering its own articulation parameters. The connections between them is updated based on Hebbian learning law through well designed interactions. By these mechanisms, the robot can acquire vowels of the caregiver even if they have different articulation parameters.

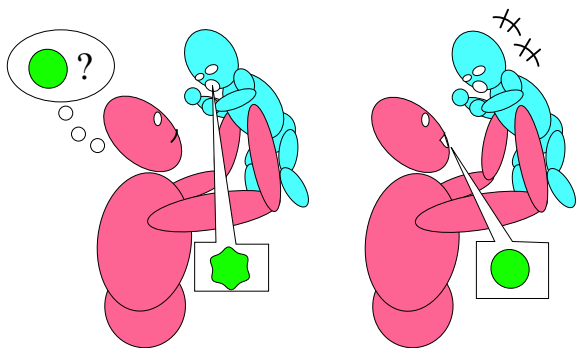
The rest of paper is organized as follows. First, we explain how interactions to learn vowels are designed. Then, we describe how the learning is proceeded with the proposed learning method. After showing the configuration of the experimental robot and preliminary experiment to confirm the acoustic property of it, we examine whether the proposed method works. Finally, we show the related work in robotics and discuss future work and give a conclusion.

2. The design of interaction

In order for a robot to learn to vocalize vowels without explicit knowledge about relation between vowels and its sensorimotor system, it should obtain information to learn by interaction with the human caregiver. Our approach is designing this interaction to work so that the learning of phonemes can be successful.

From observations in studies on infant development (Peláez-Nogueras et al., 1996, Masataka and Bloom, 1994), we conjecture that the caregiver's parrotry of infant's speech-like cooing lead the infant to acquire common phonemes with the caregiver. As a preliminary, constructive model of our conjecture, we design interaction between the robot and the caregiver by embedding a random articulation mechanism in it and by letting the caregiver imitate its articulated sound. The designed interaction follows a following process which seems to reproduce infant-mother one (see Fig. 1). Firstly, the robot articulates at random to generate vocalization in both speech-like and unspeech-like manner. If the caregiver interprets its vocalization as one of vowels which he/she usually uses, he/she utters the corresponding one.

Through such caregiver-robot interactions designed in such a way, it obtains the invariant pairs of its articulation and the corresponding vowel in the caregiver's language system so that it can succeed in the vowel acquisition with a simple learning law in spite of difference of the articulation parameters.



(a) An infant vocalizes randomly.

(b) A mother replies in the parrotry like manner.

Figure 1: A mother-infant interaction.

3. Embedded mechanism

In this section, we describe the embedded mechanism in the robot which consists of two layers and connections between them (see Fig. 2). After describing about processing in two layers, we give two learning laws of the connections between them.

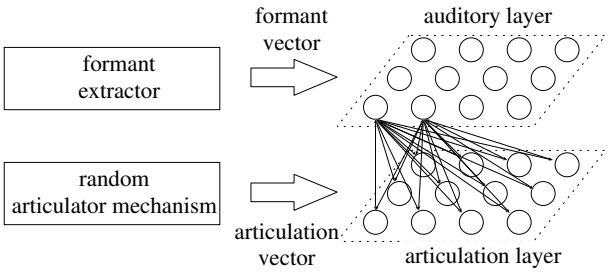


Figure 2: Learning mechanism by an auditory layer, an articulation layer and connections between them.

3.1 Auditory layer

Frequency peaks in the sound wave of vowels are effective features to distinguish vowels and called formant. The auditory layer receives *formant vectors* from the extractor of formant, that consists of the lowest four formants of the caregivers' utterances and clusters them in a self-organizing manner. For the sake of self-organizing clustering, the method of Kohonen map (Kohonen, 1984) is applied.

Let $\boldsymbol{f}_i = [f_{i1}, \cdots, f_{i4}] \in \Re^4, i = 1, \cdots, N_f$ be a code vector of the i-th unit in the auditory layer

which consists of N_f units and $\boldsymbol{r}_i^f \in \Re^2$ be a topology vector of the same unit. When it receives a formant vector $\boldsymbol{f} \in \Re^4$, units which have closer code vectors are activated more. Then, the most activated units suppresses other units. Finally, an activation a_i^f of i-th unit is calculated by

$$a_i^f = \begin{cases} g(\boldsymbol{f}_i^T \boldsymbol{f} - h) & \text{if} \quad i = \arg_i \max \boldsymbol{f}_i^T \boldsymbol{f}, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where $g(x)$ is a step function of scalar x and h is a scalar threshold. The most activated unit is called winner unit and labeled win^f.

In the algorithm of Kohonen map, code vectors of some units near the winner is modified to be close to the input vector. The updating rule is described as,

$$\begin{aligned} \boldsymbol{f}_i(t) &= \boldsymbol{f}_i(t-1) + \alpha(t) \cdot \\ & \quad \Phi(\boldsymbol{r}_i^f, \boldsymbol{r}_{win^f}^f)(\boldsymbol{f}(t) - \boldsymbol{f}_i(t-1)), \end{aligned} \tag{2}$$

where $\alpha(t)$ is a time dependent scalar learning rate and $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is a monotonic decreasing function with respect to a distance between vector $\boldsymbol{x}$ and $\boldsymbol{y}$ which is calculated by

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{y}|}{2\sigma^2(t)}\right), \tag{3}$$

where $\sigma(t)$ is a time dependent scalar which determines how near units learn. At the beginning of learning, $\sigma(t)$ is such a high value that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is high in the wide region of the auditory layer and gradually decreases so that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ of only neighborhood units have high values at the end of learning. After learning based on this rule, code vectors close to frequently observed input vectors are clustered in the auditory layer, in this case, frequently heard formants, that is vowels.

3.2 Articulation layer

The articulation layer receives *articulation vectors* from the random articulation mechanism, that consists of five motor commands to articulate by deformation of the silicon vocal tract and clusters them by the method of Kohonen map as well as the auditory layer.

Let $\boldsymbol{m}_i \in \Re^5, i = 1, \cdots, N_m$ be the i-th code vector of the articulation layer which consists of N_m units and $\boldsymbol{r}_i^m \in \Re^2$ be the i-th topology vector in it. When it receives an articulation vector $\boldsymbol{m} \in \Re^5$, an activation a_i^m of i-th unit is calculated by the same manner in eq. (1). Updating code vectors is also done by the same way in eq. (2). The most activated unit is called winner unit and labeled win^m.

3.3 Learning connections

The connections between the auditory layer and the articulation layer are updated based on Hebbian law which is a learning model of neural network. Based on this learning model, connections between neurons activated at the same time are increased while others are decreased. Let w_{ij} be a connection weight between i-th unit in the auditory layer and j-th unit in the articulation layer. The learning law is described as

$$\tau \dot{w}_{ij} = -w_{ij} + c a_i^f a_j^m, \tag{4}$$

where τ is a time constant of learning and c is a learning rate. Based on eq. (4), w_{ij} will converge to

$$w_{ij} = cE\{a_i^f a_j^m\}, \tag{5}$$

where $E\{a_i^f a_j^m\}$ is the average of $a_i^f a_j^m$ (Amari, 1977) . Actually, we use the discretizing version of the updating law (eq. (4)) such as,

$$w_{ij}(t+1) = w_{ij}(t) + \frac{1}{\tau}(c a_i^f(t) a_j^m(t) - w_{ij}(t)), \tag{6}$$

where t denotes the time stamp.

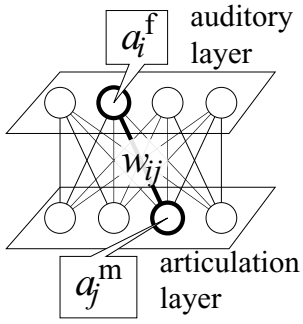


Figure 3: Connections between the auditory layer and the articulation one.

Since interaction between the caregiver and the robot is designed by the random articulation mechanism and parrot like teaching, the invariant pairs of units in the both layer are activated at the same time. Therefore, through learning process, articulations are matched with corresponding vowels as a connection between both layers.

However, such interactions may connect multiple articulation units with a corresponding vowel since the caregiver would interpret some vocalizations by different articulations as the same vowel. In order to match a heard vowel with an unique articulation to vocalize it, a modified learning law with considering facility of articulation is introduced, that is, more (less) facile articulations are increased (decreased) more. Therefore, learning law of connections are slightly modified such as:

$$w_{ij}(t+1) = w_{ij}(t) + \frac{1}{\tau}(c\eta(\boldsymbol{m}) a_i^f(t) a_j^m(t) - w_{ij}(t)), \tag{7}$$

where $\eta(\boldsymbol{m})$ is a facility function of articulation that evaluates necessary torque and intensity of deformation change calculated by

$$\eta(\boldsymbol{m}) = k \exp\left(-\frac{C_{trq}(\boldsymbol{m})}{\sigma_t^2}\right) \cdot \exp\left(-\frac{C_{idf}(\boldsymbol{m})}{\sigma_d^2}\right), \tag{8}$$

where k, σ_t, σ_d are scalar constants, and C_{trq} and C_{idf} are cost functions of torque and intensity of deformation change, respectively. σ_t and σ_d are chosen by trial and error. The cost functions are defined as

$$\begin{aligned} C_{trq}(\boldsymbol{x}) &= \boldsymbol{x}^T \boldsymbol{x}, and \\ C_{idc}(\boldsymbol{x}) &= \sum_{k=1}^{4} (x_k - x_{k+1})^2, \end{aligned} \tag{9}$$

where x_k is k-th element of the vector $\boldsymbol{x}$.

4. A test-bed robot

Vocalization is commonly regarded as a result from a modulation of a source of sound energy by a filter function determined by the shape of the vocal tract, that is often referred to as the "source-filter theory of speech production" (Rubin and Vatikiotis-Bateson, 1998). We implement the source-filter theory by using a vibrator as a sound source and silicon rubber tube as a vocal tract which can be deformed by five electric motors. This implementation is similar with one of Higashimoto et al. (Higashimoto and Sawada, 2002) in which an artificial vocal code is implemented as a sound source.

Figures 4 and 5 show an appearance and an overview of the experimental robot. Five electric motors are bound to respective attachments with piano wires and pull them to deform the silicon vocal tract. The pulling points are determined by trial and error. They are controlled by motor controller (usbMC01, iXs Research Corp.)/drivers (iMDs03, iXs Research Corp.) according to the commands from the host computer. An artificial larynx (MYVOICE, SECOM MEDICAL SYSTEM Co. Ltd.) is used as a sound source which generates a sound by vibration. Host computer receives signals from an microphone and calculates formants of them.

4.1 Preliminary experiment

We conduct a preliminary experiment to confirm the acoustic property of the silicon vocal tract in the experimental robot. The robot deforms its vocal tract, vibrates the artificial larynx and calculates formant of its vocalizations which is interpreted as Japanese vowels by the experimenter. Figure 6 shows a distributions of the calculated formants of the vocalization by the robot in which ones by human experimenter are also shown to compare. Table 1 shows averages

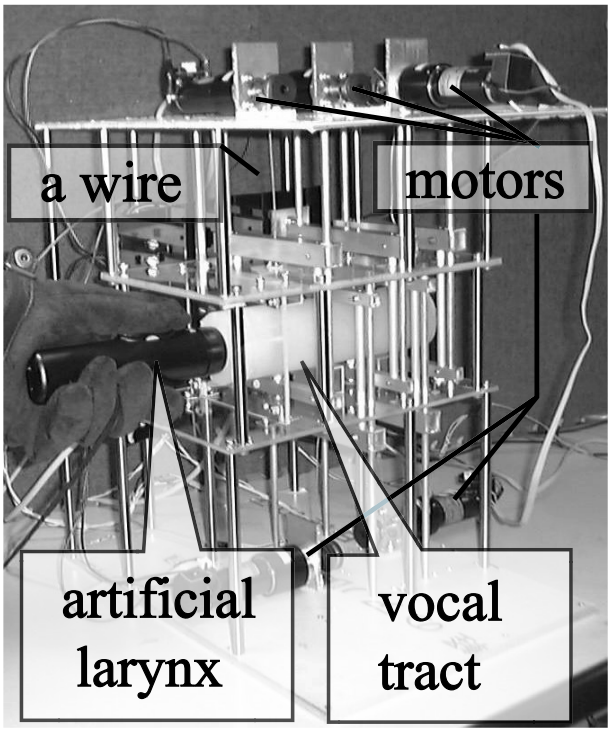


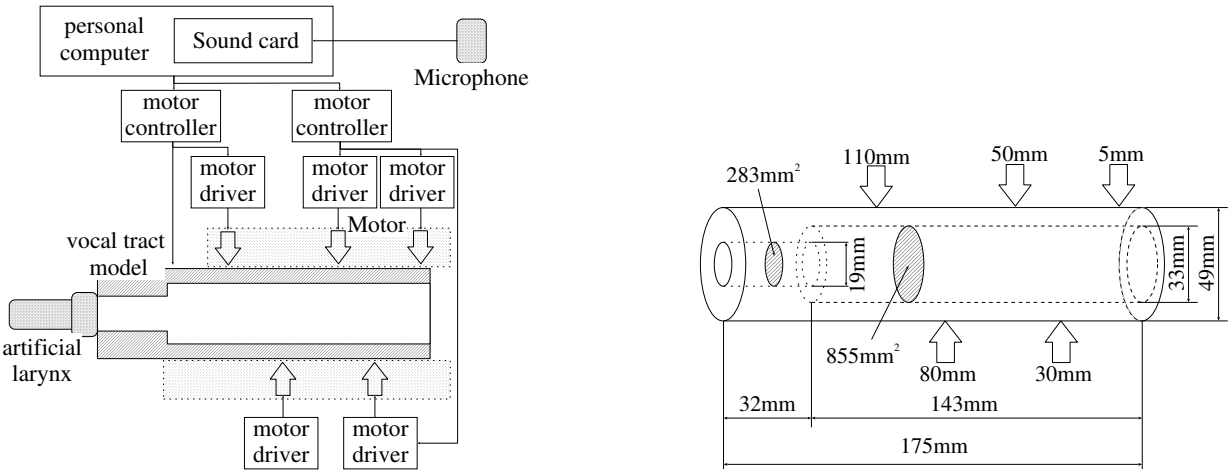
Figure 4: An appearance of the experimental robot.

of respective vowels by the robot in which the value in the bracket shows one by the human experimenter.

We confirmed that it can vocalize four Japanese vowels excluding a vowel /o/. Its distribution of formants tend to be higher than one of the human experimenter but they are clustered in the space of formants (see Fig. 6). It means that formants are available for recognizing vowels of the robot as well as ones of human beings. The reason why the robot failed to vocalize the vowel /o/ seems that the configuration of the experimental robot does not have sufficient degrees of freedom to vocalize it. In the following experiments, therefore, vowels indicate ones excluding the vowel /o/.

Table 1: Averages and standard deviation of formants of the robot and the caregiver (inside brackets) for individual Japanese vowels, /a/, /i/, /u/, and /e/.

	1st [kHz]	2nd [kHz]	3rd [kHz]
/a/	1.33 ± 0.10 (0.40 ±0.05)	1.87 ± 0.15 (0.92 ±0.04)	2.64 ± 0.13 (2.93 ±0.08)
/i/	1.23 ±0.12 (0.23±0.02)	2.02 ±0.05 (1.77±0.28)	2.83±0.06 (3.11±0.09)
/u/	0.90 ±0.70 (0.25±0.03)	2.12±0.54 (0.96±0.10)	3.21 ±0.43 (2.78±0.21)
/e/	1.46±0.05 (0.31±0.11)	1.79±0.16 (1.24±0.38)	2.73±0.07 (2.37±0.13)



(a) An overview

(b) Pulling points

Figure 5: An overview of the system and pulling points of the vocal tract.

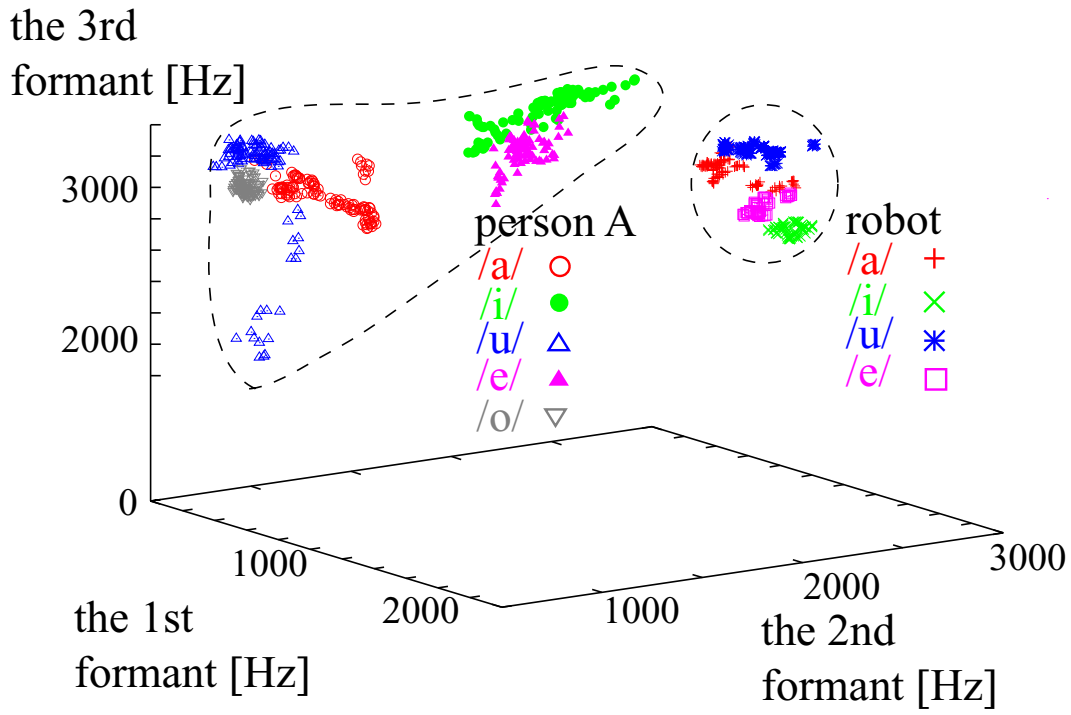


Figure 6: Formant distributions of the robot and the caregiver(person A).

5. Experiment

We conduct an examination with the experimental robot to confirm whether the proposed method works well. In this experiment, after the robot vocalizes by a random articulation mechanism, the human caregiver judges which vowel corresponds to its vocalization and utters the corresponding ones. It calculates formants of the caregiver's utterance and updates code vectors and connections according to the proposed method. In order to reduce the learning time, it corrects the data experienced in the interaction and learns by using it in the iterative computation. Each element of code vectors in the articulation layer and commands of the random articulation mechanism are quantized into five levels.

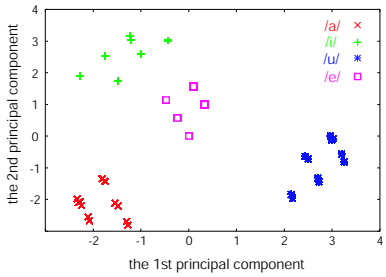
5.1 Learning without the facility criterion

Fig. 7 shows a learning result without considering the facility criterion, that is based on eq. (6). Both figures, (a) and (b), are distributions of the articulation vectors compressed into two dimensional space by a method of principal component analysis. Fig. 7(a) shows distributions of the articulation vectors in vocalizing randomly which can be interpreted as Japanese vowels by the caregiver. It is calculated which units in the articulation layer is most strongly connected with ones in the auditory layer which are activated by the caregiver's utterance of vowels and shown in Fig. 7(b).

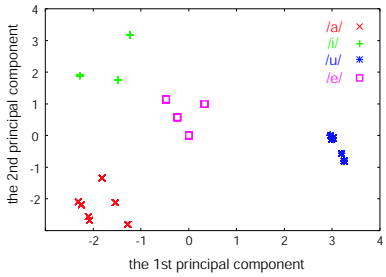
We can see that connected articulation vectors are parts of the region in which the caregiver can interpret them as vowels corresponding to his/her utterances. Actually, vocalizations by the articulation vector which is calculated by learned map are clear enough to interpret as vowels. Therefore, it is confirmed that learning vowels succeeded in by the proposed method.

5.2 Learning with the facility criterion

A learning result with the facility criterion, that is based on eq. (7), is shown in Fig. 8 in the same manner as in the Fig. 7. We can see that less articulation vectors are selected with respect to the caregiver's utterances than in Fig. 7(b). Therefore, it is confirmed that the facility criterion works so that a sound of vowel is almost matched with an unique articulation to vocalize. Remaining articulation vectors are facile to articulate.



(a) Generated by the random articulation mechanism



(b) Calculated by the acquired map for heard vowels of the caregiver

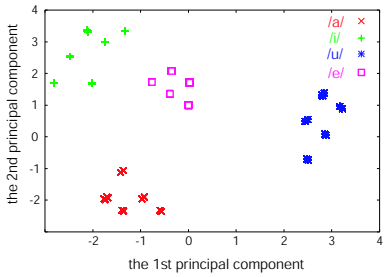
Figure 7: Learning result without the facile criterion: articulation vector distributions.

6. Related works

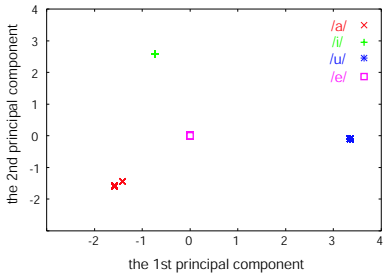
There are some related studies to acquire phonemes in robotics. Nishikawa et al. (Nishikawa et al., 2002) built an anthropomorphic robot which can produce Japanese phonemes including consonant sounds. However, articulations of phonemes were done by human manual tuning since they did not coped with acquisition problem itself. The studies based on the voice recognition and synthesis systems, for example (Kanda et al., 2002), also have not coped with it.

Higashimoto and Sawada (Higashimoto and Sawada, 2002) built a system which learns to vocalize human vowels. It consists of a vibrator which generates a source sound and a deformable silicone rubber tube which works as a filter to change the spectrum envelop of the source sound. Its vocalization of human vowels is performed by learning an inverse model of articulation parameters with respect to a spectrum envelop of sound. Therefore, it performs a kind of imitation in which it modified the generated sound toward the reference vowel sound uttered by a human experimenter based on the criterion of raw sound wave similarity. However, a human infant can imitate adult phonemes in spite that it can not vocalize the same sounds from a viewpoint of raw sound wave similarity since its vocal mechanism is under their development.

Unlike the previous studies, we argue how a robot can acquire phonemes through interactions with its caregiver since imitation in the criterion of raw sound wave does not explain the phoneme acquisition process for human beings.



(a) Generated by the random articulation mechanism



(b) Calculated by the acquired map for heard vowels of the caregiver

Figure 8: Learning result with the facile criterion: articulation vector distributions.

7. Discussion and Conclusion

In this paper, we propose a constructive model of the vowel acquisition process between agents with different articulation parameters without any explicit knowledge. In order to acquire vowels, the caregiver-robot interaction is designed by embedding a random articulation mechanism and adopting parrot like teaching based on an observation of mother-infant interactions. A validity of the proposed model is confirmed by using an experimental robot.

When an agent tries to imitate behaviors of another agent with different body structure, it needs to abstract observed behavior to some extent since it cannot perform it as they are. However, abstraction brings arbitrariness into imitation process even if the agent acquires pairs of observed behaviors and its own ones which are considered to be corresponding as in our first experiment. We proposed the method to cope with the arbitrariness by introducing a subjective criterion, that is facility to behave. As the criterion of the facility of vocalization reduces arbitrariness of matched articulations for observed vowels in our second experiment, such a subjective criterion could take an important role in imitation, understanding other's behavior, and any kinds of communicative processes between agents with different body structures. In other words, maybe needless to say, the agent including human beings seems able to perform communication only by presuming in its egocentric viewpoint.

It is speculated that an infant can adjust the pattern of its vocalizations in reaction to its mother's pattern of response from observation of contingent and non-contingent infant-mother interactions (Masataka, 1993). Although we need more observations about how they react or how the reactions are changed under their development, infant's reactions seem to play important roles in phoneme acquisition process and emergence of communication. However, our preliminary model cannot explain how infant's reactions play such roles since we apply a random articulation mechanism. Modifying it to develop to be controlled by the auditory layer can be regarded as the first step to model it.

Acknowledgement

This study was partially supported by "the Advanced and Innovational Research program in Life Sciences" and "Priority Assistance of the Formation of Worldwide Renowned Centers of Research - The 21st Century COE Program (Project: Center of Excellence for Advanced Structural and Functional Materials Design)" both of which is from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. It was also partially supported by Research Fellowships for Young Scientists from Japan Society for the Promotion of Science.

We also thank Mr. Tsukinoki for his patient support in building our experimental robot.

References

Amari, S. (1977). Neural theory of association and concept-formation. *Biological Cybernetics*, 26:175–185.

Higashimoto, T. and Sawada, H. (2002). Speech production by a mechanical model construction of a vocal tract and its control by neural network. In *Proc. of the 2002 IEEE Intl. Conf. on Robotics & Automation*, pages 3858–3863.

Kanda, T., Ishiguro, H., Imai, M., Ono, T., and Mase, K. (2002). A constrcutive approach for developing interactive humanoid robots. In *Proc. of IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*.

Kohonen, T. (1984). *Self-Organization and Assosiative Memory*. Springer-Verlag.

Masataka, N. (1993). Effects of contingent and non-contingent maternal stimulation of vocal behav-

ior of 3- to 4-month-old japanese infants. *J. of Child Language*, 20.

Masataka, N. and Bloom, K. (1994). Accoustic properties that determine adult's preference for 3-month-old infant vocalization. *Infant Behavior and Development*, 17:461–464.

Nishikawa, K., Imai, A., Ogawara, T., Takanobu, H., Mochida, T., and Takanishi, A. (2002). Speech planning of an anthropomorphic talking robot for consonant sounds production. In *Proc. of the 2002 IEEE Intl. Conf. on Robotics & Automation*, pages 1830–1835.

Peláez-Nogueras, M., Gewirtz, J. L., and Markham, M. M. (1996). Infant vocalizations are confitioned both by maternal imitation and motherese speech. *Infant behavior and development*, 19:670.

Rubin, P. and Vatikiotis-Bateson, E. (1998). *Animal Acoustic Communication*, chapter 8 Measuring and modeling speech production. Springer-Verlag.