Synergistic Intelligence: Towards Emergence of Communication - A Preliminary Study on Vowel Acqusition by Maternal Imitation -

Minoru Asada, Katsushi Miura, and Yuichiro Yoshikawa

***Abstract*— This paper presents an introduction of the JST ERATO Asada Synergistic Intelligence Project (hereafter, SI) and its preliminary study on emergence of communication. The project aims at realizing a new understanding and construction of human intelligence from a complex of science and technology through the design, implementation, and operation of humanoids, and by verifying the constructive model using the methods of cognitive and brain sciences. The new domain based on this complex is designated "synergistic intelligence." This name refers to both the co-creation of intelligence by science and technology, and the co-creation of intelligence through the interactions between the body and environments.**

As a preliminary study, vowel acquisition based on visual and auditory mutual imitation is given. It aims at modelling mother-infant interaction in vowel acquisition based on the findings in infant studies, especially focusing on the roles of maternal imitation (i.e., imitation of the robot voices by the caregiver) since it could play a role to instruct the correspondence of the sounds. Furthermore, we suppose that it causes *unconscious anchoring* in which the imitated voice by the caregiver is performed unconsciously, closely to one of his/her own vowels, and hereby helps to leading robot's utterances to be more vowel-like. Through the experiments with a Japanese imitative caregiver, we show that a robot succeeds in acquiring more vowel-like utterances than one without such a caregiver.

I. INTRODUCTION

The twenty-first century is known as the "Century of the Brain" and will be an era of robot habitation with humans. Currently, these notions are not related to one another, despite the related subjects of understanding and realizing the emergence of human intelligence. With advanced technologies, such as noninvasive imaging devices, recent advances in neuroscience are now approaching problems regarding cognition and consciousness, which have not yet been dealt with in science. However, current means are not sufficient to proceed with research regarding communication and language acquisition based on "embodiment" that epitomize the abilities of cognition and development. It is important for life sciences in the 21st century to regard the brain as a complete system, and to clarify how brains realize function through interactions with the external world based on its dynamics.

Humanoids are a class of human-type robot, and Japan is a world leader in this field. Despite rapid technological development, only superficial functions have been realized and there is no established design methodology for the emergence of human intelligence based on embodiment. It is indispensable to merge the science and technology through design, implementation, and operation of artifacts (robots) after clarifying the background deep-layer structures for the emergence of intelligence based on not simply the engineering to realize superficial functions but also the collaboration with the science fields, such as neuroscience and cognitive sciences. If these fields are merged organically, a means of verification utilizing robot technologies can be expected to promote a new brain science unique to Japan. We may also synthetically tackle the subjects of consciousness and mind, which used to be difficult to understand with conventional philosophy, psychology, sociology, cognitive science, linguistics and anthropology. Designing artifacts capable of passing these verifications necessitates innovations in current materials, such as robot sensors and actuators, and conventional artificial intelligence and control technologies.

Based on this basic concept, this study derives the design of intelligence lacking in current humanoid studies from interactions with a dynamic environment, including humans (interactions between environment and robot or between human and robot). In other words, we realize a new understanding and construction of human intelligence from a complex of science and technology through the design, implementation, and operation of humanoids, and by verifying the constructive model using the methods of cognitive and brain sciences. The new domain based on this complex is designated "synergistic intelligence." This name refers to both the co-creation of intelligence by science and technology, and the co-creation of intelligence through the interactions between the body and environments.

This new domain has four aspects (see Fig.1). The generation of dynamic motions by artificial muscles allows co-creation of the intelligence though the interaction between the body and the environment and is called "Physically synergistic intelligence (Physio-SI)." The emergence of somatosensory and motion in fetal models allows co-creation in the uterine environment. In the process of developing cognition from various interactions with the fosterer, the fosterer is the most important environmental factor. Therefore, the synthetic model of the process is called "Interpersonally synergistic intelligence (Perso-SI)." In the emergence of communication between many humans and robots, the effect of the multi-agent environment is important and "Socially synergistic intelligence (Socio-SI)" serves as the core. "Synergistic intelligence mechanism (SI-Mechanism)" verifies these processes by comparing autism and William's syndrome, which highlight extereme aspects of language and cognition capabilities, and promote the construction of a

The authors Asada Synergistic Intelligence Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Frontier Research Center-I, Graduate School of Eng., Osaka Univ., 2-1 Yamada-oka, Suita, Osaka 565-0871 Japan {asada,miura,yoshikawa}@jeap.org

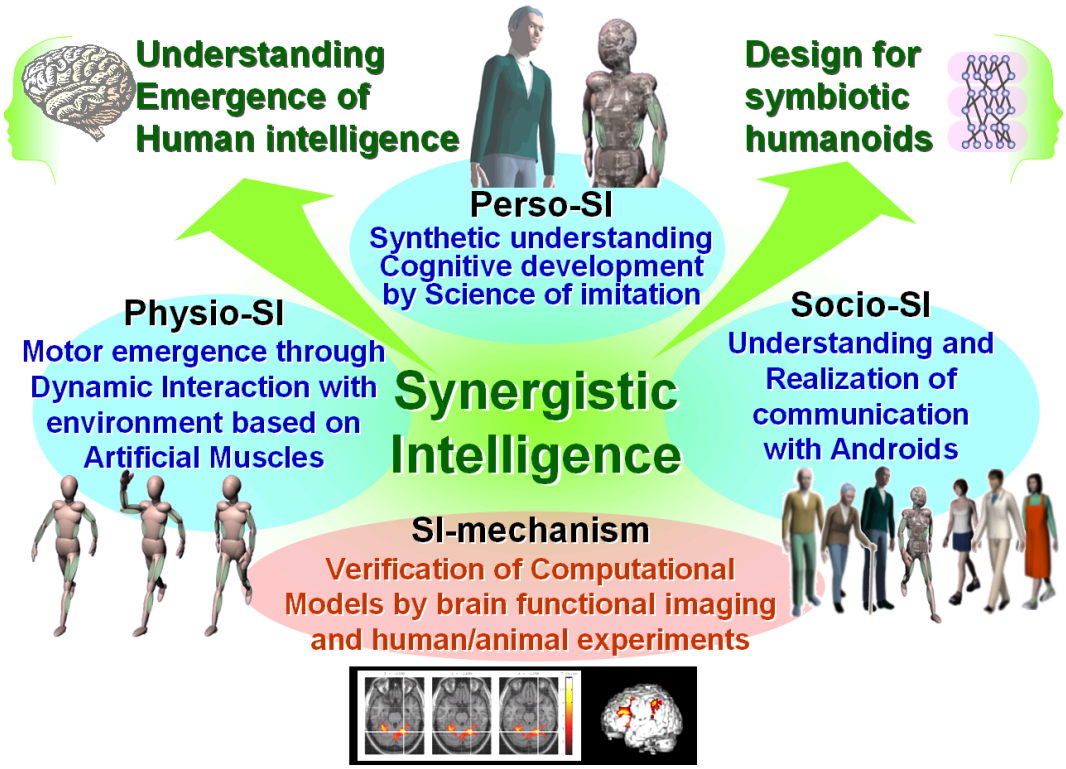


Fig. 1. An overview of Synergistic Intelligence Project

new constructive model. We will promote harmonious study within the group, and design strategies to realize a synergistic intelligence system in humanoids that require close linkage between groups.

As a preliminary study, vowel acquisition based on visual and auditory mutual imitation is given. It aims at modelling mother-infant interaction in vowel acquisition based on the findings in infant studies, especially focusing on the roles of maternal imitation (i.e., imitation of the robot voices by the caregiver) since it could play a role to instruct the correspondence of the sounds. Furthermore, we suppose that it causes *unconscious anchoring* in which the imitated voice by the caregiver is performed unconsciously, closely to one of his/her own vowels, and hereby helps to leading robot's utterances to be more vowel-like. Through the experiments with a Japanese imitative caregiver, we show that a robot succeeds in acquiring more vowel-like utterances than one without such a caregiver.

II. Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories

It is suggested that humans generally tend to anthropomorphize the artifacts [1], and such a tendency can be amplified in facing with a humanoid robot since the similarities in appearance with humans help them easily find the correspondences between a human and a robot. Therefore, in case of communication, humanoid robots are expected to communicate with humans in a natural manner that humans do such as vocal communication. However, it is a formidable issue how humanoid robots can show the behaviors to be considered as the corresponding humans' ones since the body structure is different from each other.

On the other hand, human infants seem to successfully solve the similar problem in the language acquisition process since infants cannot regenerate the caregiver's voices as they are due to the sensorimotor im-maturities, i.e., differences in body structure. During the process, imitation seems to have a very important role regardless of the body difference, and from the viewpoint of cognitive developmental robotics [2], the study of imitation between a human and a robot is expected not only to contribute to the studies on understanding infant cognitive development process but also to provide the robot behaviors based on these studies.

Using a robot that can generate vowels with an artificial vocal band and tract (ex. [3], [4]) is one approach to directly attack the problem of imitation between dissimilar bodies. With such a vocal robot, Yoshikawa et al. [5] proposed a mother-infant interaction model for infant vowel acquisition based on the observations in developmental psychology. Inspired by the findings that maternal imitation effectively reinforces infant vocalization [6] and that its speech-like cooing tends to invoke utterances by its mother [7], they have suggested that maternal imitation (i.e., imitation of the robot's utterance by the caregiver) using adult phonemes plays an important role in phoneme acquisition, namely matching its articulations and the corresponding caregiver's utterances. In their model, the robot could find a lot of

candidates of vowels but know which of them are more vowel-like, in other words which of them are easier for humans to recognize them as vowels.

We suppose that the maternal imitation could play another important role in vowel learning beyond the role of giving the instruction to match the caregiver's vowel to the robot's utterance, that is leading robot's utterance to be more vowel-like. In this paper, we presents an interaction paradigm and experiments in order to show this another role of the maternal imitation.

The test task for a vocal robot is learning how to articulate vowel-like sounds through the interaction with a caregiver who tries to imitate the robot utterances but cannot regenerate them as they are due to the difference between their articulatory systems. In this setup, it is conjectured that the imitated voice by the caregiver is performed unconsciously, closely to one of his/her own vowels, and we call such a behavior "unconscious anchoring". Maternal imitation and this unconscious anchoring would cause two phenomena that support learning of more vowel-like sounds: (1) given maternal imitation, the robot can obtain the references to modify the mapping between the sound feature vectors of vowels generated by the caregiver and that by the robot, and (2) furthermore, by unconscious anchoring, the references would be gradually shifted to more vowel-like sounds.

In the following, we introduce the idea of unconscious anchoring and a learning mechanism based on it. Through some experimental trials of vowel learning with a Japanese imitative caregiver, we show that the robot succeed in acquiring more vowel-like utterances compared to the robot utterances without such a caregiver.

A. Assumptions and basic ideas in unconscious anchoring

An interaction model between a caregiver and a robot is shown in Figure 2 where a vocal robot interacts with a caregiver through vocalization and hearing the caregiver's voices. In this scenario,

R: the robot tries to utter one of Japanese vowels, and

C: the caregiver listens to the robot's utterance, looks at the shape of robot lip, and then tries to imitate the voice of the robot.

Such a caregiver's imitation is expected to give the robot the information how the robot's voices are interpreted by the caregiver, that seems to tell one of the most important aspects to attain the communication.

The task of the robot through such interaction is learning to find the ways of articulation by which it can generate the sounds corresponding to the caregiver's vowels. The robot cannot generate the exactly same sound as the caregiver's one and vice versa since the articulatory system is different from each other. In other words, the regions of sounds that the caregiver and the robot can generate are usually different from each other or do not overlap with each other at worse. Nevertheless, humans can map the robot's sounds to their own corresponding vowels [5]. On the contrary, it is usually not trivial for the designers to provide their robots with the accurate mapping between these two regions of sounds.

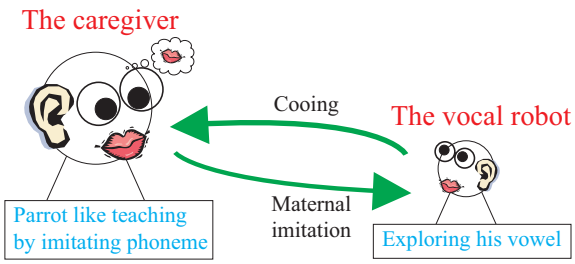


Fig. 2. An interaction model between an imitative caregiver and a vocal robot

Therefore, we assume that we can provide only the rough estimates for the mapping function.

The human's utterances can be clustered in the space of the static feature of sound-wave, namely *formant*, in which clusters correspond to vowels [8]. Therefore, it is feasible to assume that we can provide the robot with the categories of the desired vowels or that the robot learns them through the observation of human's usual utterances.

Owing to these above assumptions, when it listens to the caregiver's imitative utterance, the robot can obtain the information how its attempting voice differs from the desired vowel category. Then, it can obtain the rough information of the difference in its own regions of sounds by using the mapping function. The mapped difference can be used for modifying its own 'vowel' category. The phenomenon of leading robot's utterance to be more vowel-like would occur by virtue of the following implicit assumption underlying the mutual imitation process. While he/she attempts to imitate the robot's voice, the caregiver unconsciously uses his/her own voice and vowel due to the sensorimotor constraints. In other words, the caregiver's imitative voice is slightly biased to the direction towards his/her own vowel category. Consequently, since the directions of modifying the robot's categories are biased towards the ones corresponding to the caregiver's vowels, the robot voices would gradually become more vowel-like, i.e., easier for humans to recognize them as vowels.

B. Learning method

The robot learns how to articulate the vowels corresponding to the caregiver's ones through mutual imitation. In the learning process, the 'vowel' categories of the robot defined in the *formant space* are updated through the interaction with an imitative caregiver. In this subsection, we introduce how we provide the robot with rough estimation of the mapping by which it can convert the information of the correspondence onto the region of its own generable sound. We then introduce the updating rule of the 'vowel' categories of the robot.

1) Mapping functions between the regions of generable sounds: Human's vowels are well distinguished in the formant space, a well-known sound feature space for vowel classification [8]. Figure 3 shows sample distributions of five Japanese vowels uttered by a Japanese male and a Japanese

female. As you can see from Figure 3, the categories of Japanese vowels are distributed in the formant space as if they form a pentagon.

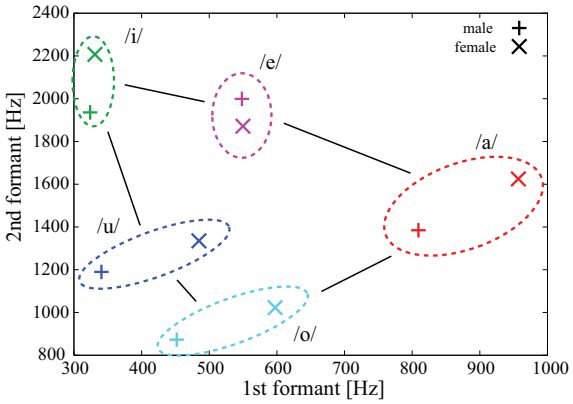


Fig. 3. A sample distribution of human vowels in the formant space

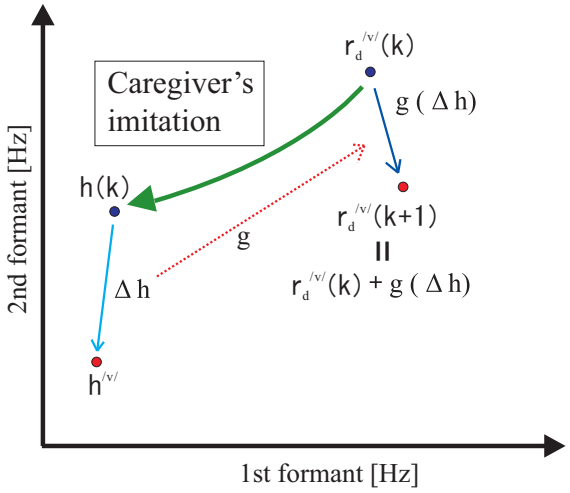


Fig. 4. Updating process of the prototype vector of a vowel ($/v/$) category $\mathbf{r}_d^{/v/}$

Since we suppose that forming a pentagon in the formant space is an important feature for vowel categories, possible pentagons in the regions of generable sounds by the robot are expected to be feasible starting positions for learning. Therefore, we provide the robot with a linear transformation as a mapping function from the region of generable sounds by the caregiver to one by the robot. In other words, the sound of the caregiver's vowel $\mathbf{h}^{/v/}$ ($/v/ = /a/, /i/, /u/, /e/$, or $/o/$) is converted to the sound corresponding $\mathbf{h}'^{/v/}$ by a mapping function $\mathbf{g}$ with the parameters of a scaling coefficient α, a rotational matrix $\mathbf{R}(\theta)$ by the angle θ, and an offset vector $\mathbf{s}$ such as

$$\mathbf{h}'^{/v/} = \mathbf{g}(\mathbf{h}^{/v/}; \alpha, \theta, \mathbf{s}) \equiv \mathbf{r}_c + \alpha \mathbf{R}(\theta)(\mathbf{h}^{/v/} - \mathbf{h}_c) + \mathbf{s} \quad (1)$$

where $\mathbf{h}_c$ and $\mathbf{r}_c$ indicate the centroids of respectively.

2) Updating the 'vowel' categories of the robot: The imitated voice of the robot utterance by the caregiver is supposed to tell the difference of the robot utterance from the sound of the closest vowel category of the caregiver. The differences can be converted to the ones by the robot based on the mapping function and be used to update the 'vowel' categories of the robot.

Suppose that the robot utters $\mathbf{r}_d^{/v/}$ that is one of the current prototype vowel category of $/v/$ and the caregiver generates the imitated sound $\mathbf{h}$. Let the prototype category of the usual caregiver's vowel $/v/$ be $\mathbf{h}^{/v/}$. The robot updates $\mathbf{r}_d^{/v/}$ based on the difference between $\mathbf{h}$ and $\mathbf{h}^{/v/}$.

C. Experiment

In the experiments, we like to verify our hypotheses on the role of maternal imitation in the acquisition process of more vowel-like utterances by the robot: (1) the imitated voices by the caregiver converge on his/her own vowels owing to "unconscious anchoring" regardless of different mapping functions, (2) the vowels that the robot acquired through the maternal imitation are more acceptable as Japanese vowels than ones acquired from the fixed desired formant vectors.

We used four types of rough estimation in the experiments: (a) translation to match the centroids, (b) translation plus scaling, (c) translation plus offset, and (d) translation plus rotation.

First, we show our vocal robot and how it utters. Next, the experimental procedures are explained, and then the results on the imitated and acquired vowels with statistical analysis are given.

1) The vocal robot: Vocalization is commonly regarded as the result from a modulation of a source of sound energy by a filter function determined by the shape of the vocal tract; this is often referred to "source-filter theory of speech production" [9] and implemented also in the previous studies [5], [4]. To model the process of vowel convergence in mother-infant interaction, we improved the vocal robot used in previous study [5] in such a way that we replaced the sound source with an air compressor and an artificial vocal band, and added a lip at the front end of the vocal tract, and the length of the robot's vocal tract changes from 170 [mm] (male's average vocal tract length) to 116 [mm].

Figure 5 shows the new vocal robot. The compressed air is conveyed through a tube to the artificial vocal band to generate the source sound of fundamental frequency, then the sound-wave is spread out through the vocal tract and the lip, that is a silicon tube with hollow end which resembles a human lip. To modulate the sound-wave, the vocal tract and lip were wired with four electric motors, respectively, and could be deformed by them. The host computer controls the motors through the motor controllers (usbMC01, iXs Research Corp.). The host computer receives signals from a microphone and calculates their formants.

The vocal robot has six degrees of freedom, two of which are used for opening/closing of lips by four motors, and four of which for deforming the vocal tract by another set of motors. First, we show the utterance capability of the robot. The motor commands which control the shape of vocal tract are quantized into five levels, 0 (free, no deformation), 0.25,

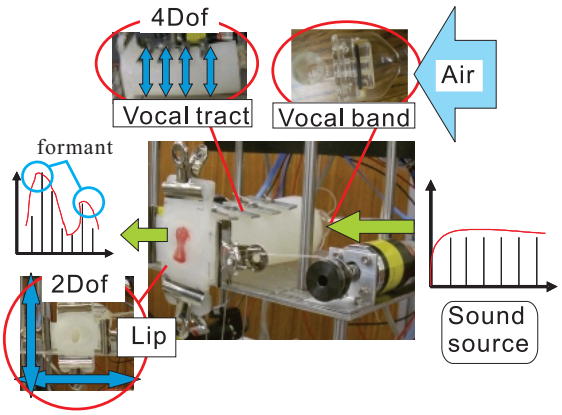


Fig. 5. The articulatory system of the vocal robot

0.5 (medium), 0.75, and 1.0 (maximum deformation), and the motor commands which control lip shape are assigned to imitate the shape of human lips. Table I shows the motor commands used to imitate human lip shape, and Figure 6 shows the formant distribution of the robot utterances in the formant space where the horizontal and vertical axes indicate the first and second formants, respectively. Furthermore, Figures 7 (a), ···, and (e) show the formant distributions categorized by the lip shape. In Figure 7 formant distribution relates with the robot's lip shape, and the larger opening size is, the higher the first and second formant shift.

TABLE I

THE MOTOR COMMANDS TO FORM THE LIP SHAPES TO RESEMBLE HUMAN'S ONES IN VOCALIZING VOWELS

Motor output	/a/	/i/	/u/	/e/	/o/
vertical direction	1.0	0.0	0.0	0.5	0.5
horizontal direction	1.0	1.0	0.0	1.0	0.0

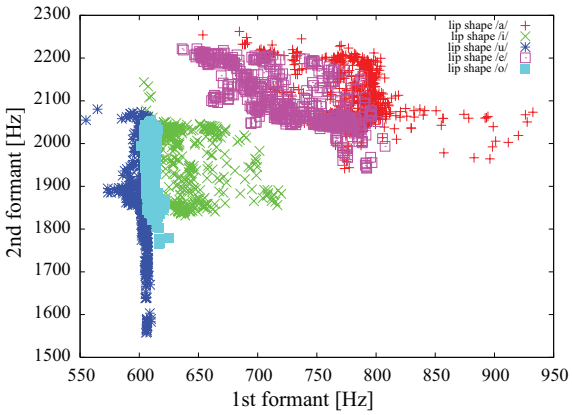
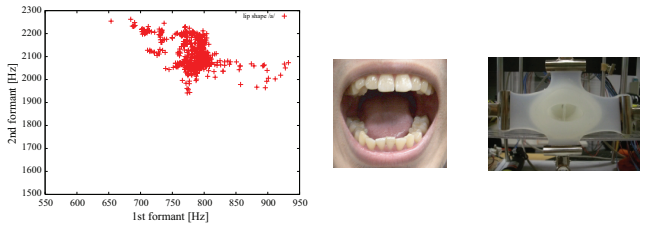
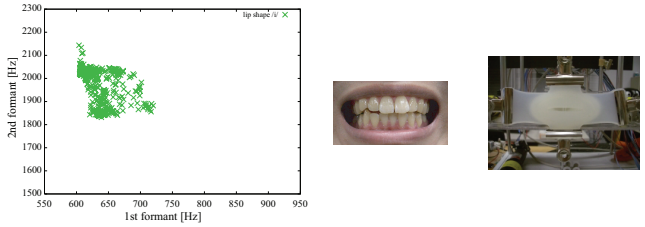


Fig. 6. The distribution of the robot utterances in the formant space

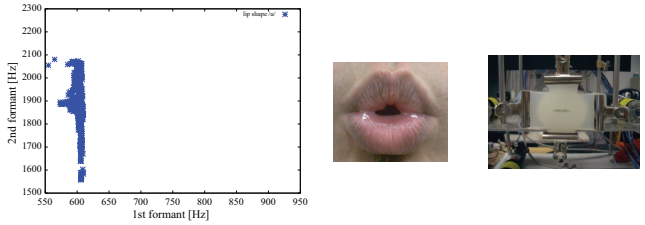
By using the data in Figure 6 as the list of the pairs of the motor commands and the formant vectors, the robot can generate the desired sound. From the list of the pairs, it can finds some candidate pairs of which formant vectors are



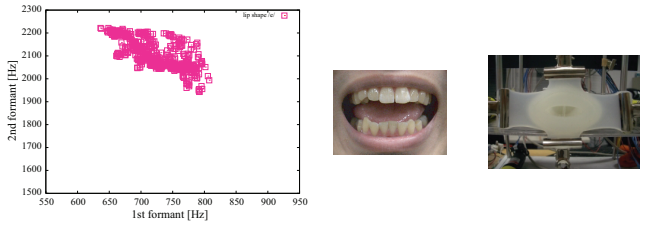
(a) Lip shape /a/



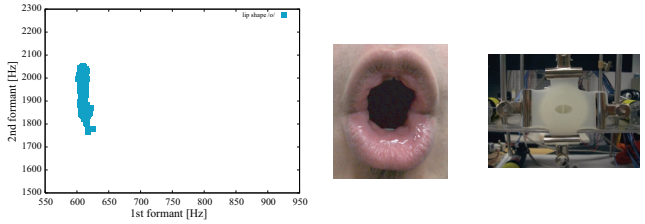
(b) Lip shape /i/



(c) Lip shape /u/



(d) Lip shape /e/



(e) Lip shape /o/

Fig. 7. The distributions of the robot utterances in the formant space each of which is generated with a lip shape that resembles human's one for the corresponding vowel utterance

sufficiently close to the desired one. Then, it selects a pair from the candidates, which has the closest motor command to the previous motor one.

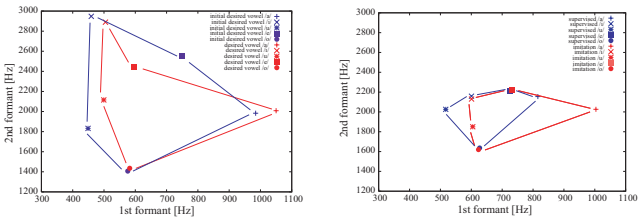
2) Set up and procedure: The experiments are conducted on the condition where one subject (the same caregiver through the all experiments) participated in the vowel acquisition process by two kinds of methods with four kinds of mapping functions each, that is, totally eight kinds of experiments to verify the hypotheses. Note that each experiment is iterated five times for the later statistical analysis. In the vowel acquisition process with the proposed method of maternal imitation, the caregiver tries to imitate the robot's utterances as other person would judge his imitated voice is the same as the robot's one. Through the turn taking of uttering voice, the robot modifies the desired formant vectors by using the caregiver's utterances as the information of the correspondence of both utterances. For the comparison, the other process of vowel acquisition is performed by a supervised learning method with fixed desired formant vectors specified by the mapping function. The variations of four kinds of mapping functions are as follows:

- translation: only translation by the difference between two centroids: α=1.0, **R**(0), **s**=(0, 0).
- scaling: translation plus scaling: α=0.24 (that coincides with the region of the generable sounds of the robot), **R**(0), **s**=(0, 0).
- offset: translation plus offset: α=1.0, **R**(0), **s**=(-100, 200).
- rotation: translation plus rotation: α=1.0, **R**(30), **s**=(0, 0).

The number of steps for the supervised learning and the number of turn takings for the maternal imitation are 20 for each vowel category.

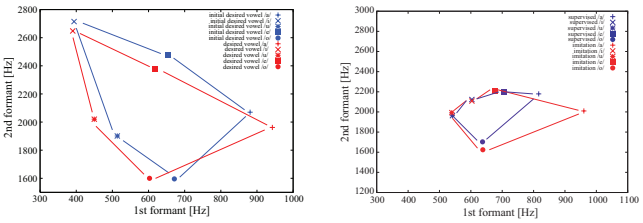
3) Results: First, we present the robot's vowels that acquired through experiments. Figure 8 shows the vowel categories in the formant space that the robot acquired with the supervised learning and the maternal imitation with a mapping function (translation). In Figure 8 (a), the desired formant vectors in the supervised learning and the final desired formant vectors modified in the proposed learning process with the maternal imitation are indicated as blue symbols (+, * etc.) and red ones. Hereafter, blue and red colors indicate the data by the supervised learning and the maternal imitation, respectively. In Figure 8 (b), the vowel categories as formant vectors acquired by the both methods are indicated in the same colors as Figure 8 (a). Figures 9 (a) and (b) show the similar graphs as Figures 8 (a) and (b) in the case of averaging the vowel categories among four mapping functions. The differences between the supervised learning and the learning with maternal imitation in Figure 8 and Figure 9 imply that the robot succeeded in modifying its desired formant vectors.

We hypothesized the unconscious anchoring that gradually leads the caregiver's utterance to his/her own vowels, and to check this tendency, the changes of the difference $\Delta\mathbf{h}$ in Figure 4 (the distance indicating the error of the mapping) at the beginning and at the end of the learning are examined.



(a) The desired formant vectors (b) The acquired formant vectors

Fig. 8. The vowel categories in the formant space that the robot acquired by the supervised learning and the maternal imitation with a mapping function (translation)



(a) The desired formant vectors (b) The acquired formant vectors

Fig. 9. The average vowel categories acquired by the supervised learning and the maternal imitation

This change is shown in Figure 10 where the vertical axis indicates the size of $\Delta\mathbf{h}$ and the vertical bars indicate average of first three times learning of ones and average of last three times learning of them that acquired through five times experiment with each mapping functions and narrow bars indicate standard deviation of them. From the T-test, there appeared to be highly significant difference of the average size of $\Delta\mathbf{h}$ between in the first three steps and in the last three steps ($p = 2.0 \times 10^{-5}$). This difference implies the tendency of the convergence of the caregiver's imitation, hopefully to his/her own vowels. This result seems to support the verification of the first hypothesis.

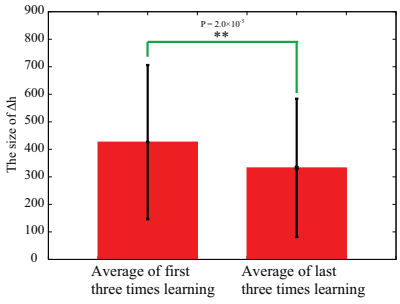


Fig. 10. The difference between the imitated voices by the caregiver and his/her corresponding usual vowels at the beginning and at the end of the interaction

Since it is difficult to show which is more vowel-like between two methods (the maternal imitation and the supervised learning) in the formant space, we are now preparing the experiments of subjective criterion to judge it.

III. Conclusion

The brief introdcution of SI project is given and one preliminary study was shown toward emergence of communication. More studies on body representation, imitation, and symbol emergence will be done in our project and hopefully, we like to build a humanoid that can show intelligent, adaptive behaviors including dynamic motions such as runnig and junmping, conversation skill though poor utterance, and facial expression to communicate with us.

IV. Acknowledgments

The first author appreciates the group leaders of the project: Profs. Hiroshi Ishiguro and Koh Hosoda at Osaka University, Prof. Toshio Inui at Kyoto University, and Prof. Yasuo Kuniyoshi at University of Tokyo for the invaluable discussion with them.

References

[1] Byron Reeves and Clifford Nass. *The media equation -how people treat computers, television, and new media like real people and places*. Stanford Univ Center for the Study, 1996.

[2] 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*, 37:185–193, 2001.

[3] Kotaro Fukui, Kazufumi Nishikawa, Shunsuke Ikeo, Eiji Shintaku, Kentaro Takada, Hideaki Takanobu, Masaaki Honda, and Atsuo Takanishi. Development of a talking robot with vocal cords and lips having human-like biological structurest. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2526–2531, Augst 2005.

[4] T. Higashimoto and H. Sawada. 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, 2002.

[5] Yuichiro Yoshikawa, Minoru Asada, Koh Hosoda, and Junpei Koga. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, 15(4):245–258, Dec 2003.

[6] 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*, 19:670, 1996.

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

[8] R. K. Potter and J. C. Steinberg. Toward the specification of speech. *Journal of the Acoustical Society of America*, 22:807–820, 1950.

[9] Philip Rubin and Eric Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter 8 Measuring and modeling speech production. Springer-Verlag, 1998.