Finding the Correspondence of Caregiver's Vowel Categories Based on Unconscious Anchoring in Maternal Imitation

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*— Due to differences in body structure between robots and humans, it is a formidable task for robots to show behaviors that correspond to human behaviors. As a simple case of this correspondence problem, this paper presents a robot that learns to vocalize vowels through interaction with its caregiver. Inspired by the findings in developmental psychology, we focus on the role of maternal imitation (i.e., imitation of a robot voice by a caregiver), which could play a role in guiding the correspondence of sounds. Furthermore, we suppose that it causes *unconscious anchoring* in which the imitated voice by the caregiver is approaching to one of his/her own vowels without his/her intension, and thereby works for guiding robot's utterances to be more vowel-like. We propose a method for vowel learning with an imitative caregiver under the assumption that the robot knows the desired categories of caregiver's vowels and the rough estimate of mapping between the region of sounds that the caregiver can generate and the region that the robot can generate. Through experiments with a Japanese imitative caregiver, we show that a robot succeeds in acquiring more vowel-like utterances than would be possible without such a caregiver, even when the robot is provided different mapping functions.**

I. INTRODUCTION

It has been suggested that humans tend to anthropomorphize objects [1], and such a tendency may be amplified for a humanoid robot because of the similarity in appearance with humans facilitates the identification of correspondences between human and robot. Therefore, in case of communication, humanoid robots are expected to communicate with humans in a natural 'human' manner. However, determination of methods that allow humanoid robots to exhibit behaviors that correspond to human behaviors is a formidable task since the body structure of a robot is different from that of a human.

On the other hand, human infants seem to successfully solve a similar problem in the language acquisition process because infants cannot perfectly regenerate the caregiver's voices due to sensorimotor immaturities, i.e., differences in body structure. During the language acquisition process, imitation seems to have a very important role, regardless of the body difference. 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 studies on understanding the infant cognitive development process but also to provide the design theory of robot behaviors based on these studies.

Learning to vocalize vowels seems like one of the simplest tasks in imitation between dissimilar bodies because an imitator can focus only on the static features in the sound waves to be generated. Learning to vocalize vowels also seems to be the first step in infant language acquisition, which begins at two or three months of age. Previous studies have elegantly demonstrated that a population of computer-simulated agents with a vocal tract and cochlea could self-organize shared vowels through imitating each other [3], [4]. However, these studies 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.

Using a robot that can generate vowels with an artificial vocal band and vocal tract (e.g. [5], [6]) is one approach to directly attack the problem of imitation between dissimilar bodies. Using such a vocal robot, Yoshikawa et al. [7] proposed a mother-infant interaction model for infant vowel acquisition based on observations in developmental psychology. Inspired by the findings that maternal imitation effectively reinforces infant vocalization [8] and that its speech-like cooing tends to invoke utterances by its mother [9], they suggested that maternal imitation (i.e., imitation of the robot's utterances by the caregiver) using adult phonemes plays an important role in phoneme acquisition, namely instructing the matching between its articulations and the corresponding caregiver's utterances. In their model, the robot was able to find many candidate vowels but was not able to determine which of the candidates were more vowel-like, i.e., which of the candidates were easier for humans to recognize as vowels.

In the present study, we examine the hypothesis that maternal imitation could play another important role in vowel learning, that is, guiding the robot's utterance to become more vowel-like. The test task for a vocal robot is learning how to articulate vowel-like sounds through interaction with a caregiver who attempts to imitate the utterances of the robot. Note that the caregiver cannot regenerate the utterances 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 to be more similar to one of his/her own vowels. This behavior is referred to as "unconscious anchoring". Maternal imitation and unconscious anchoring are thought to provide two phenomena that support learning of more vowel-like sounds: (1) given maternal imitation, the robot can obtain references based upon which to modify the mapping between the sound feature vectors of the 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.

Based on the same supposition, Miura et al. [10] reported the possibility that the robot could acquire vowel-like sounds through interaction with an imitative caregiver. It was assumed that the designer cannot provide the vocal robot with the utterances corresponding to the caregiver's natural vowels but that the designer can provide the vocal

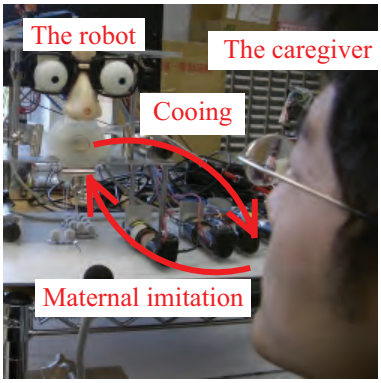


Fig. 1. Interaction model between the caregiver and the vocal robot

robot with a correct mapping between both acoustic feature spaces. However, it is generally not trivial for the designer to find such an accurate mapping for the case in which the bodies of both agents are dissimilar. In this paper, we examine another potential effect of unconscious anchoring, in which we need only obtain a rough estimate of such a mapping in order for the robot to acquire vowel-like sounds through interaction with the caregiver who shows unconscious anchoring.

In the remainder of this paper, we introduce the idea of unconscious anchoring and a learning mechanism based thereon. Through experimental trials of vowel learning with a Japanese imitative caregiver, we show that the robot is able to acquire more vowel-like utterances compared to robot utterances without such a caregiver. Furthermore, we show that the robot is able to learn vowel-like sounds even when the robot is provided different mapping functions by which to decide the correspondence between the human voice and the robot voice.

II. Assumptions and Basic Concepts in Unconscious Anchoring

An interaction model between a caregiver and a robot is shown in Fig. 1, in which a vocal robot interacts with a caregiver through vocalization and hearing the voice of the caregiver. In this scenario,

R: the robot tries to utter one of Japanese vowels, and
C: the caregiver listens to an utterance by the robot, looks at the shape of the lips of the robot, and then tries to imitate the voice of the robot.

Such imitation by a caregiver is expected to provide the robot with information as to how the voice of the robot is interpreted by the caregiver, which seems to reveal the most important aspects with regard to achieving communication.

The task of the robot through such interaction is learning to find methods of articulation by which to generate the sounds corresponding to the vowels of the caregiver. The robot cannot generate exactly the same sound as the caregiver (and vice versa) because their articulatory systems are different. In other words, the regions of sounds that the caregiver and the robot can generate are usually different from each other or do not even overlap with each other. Nevertheless, humans can map the sounds of the robot to their own corresponding vowels [7]. In contrast, it is usually not trivial for the designers to provide their robots with an accurate mapping between these two regions of sounds. Therefore, we assume that we can provide only rough estimates of the mapping function.

Human utterances can be clustered in the space of the static feature of the sound wave, namely *formant*, in which clusters correspond to vowels [11]. Therefore, it is feasible to assume that we can provide the robot with the categories of the desired vowels or that the robot will learn them through the observation of usual human utterances.

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

The idea of unconscious anchoring would be generalized to other modalities such as vision (and motion, i.e., gesture) and hopefully guide a new methodology of providing a robot with social skills through interaction. As the first step, in this paper, we focus on the issue of robot vocal acquisition through mutual imitative interaction with vision and audition.

III. Learning Method

The robot learns how to articulate the vowels corresponding to those of the caregiver through mutual imitation. In the learning process, the 'vowel' categories of the robot defined in the *formant space* are updated through interaction with an imitative caregiver. In this section, we introduce a method to provide the robot with a 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.

A. Mapping functions between the regions of generable sounds

Human vowels are well distinguished in the formant space, a well known sound feature space for vowel classification [11]. Figure 2 shows sample distributions of five Japanese vowels uttered by a Japanese male and a Japanese female. As shown in Fig. 2, the categories of Japanese vowels are distributed in the formant space as if they form a pentagon.

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

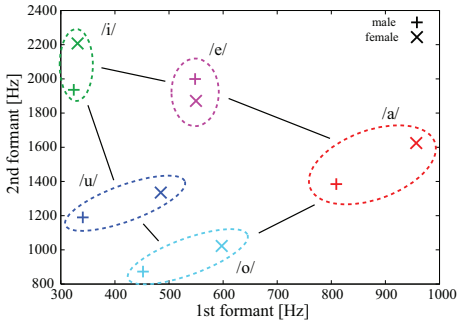


Fig. 2. Sample distribution of human vowels in the formant space

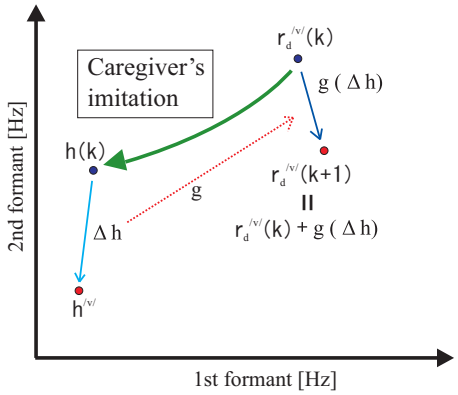


Fig. 3. Updating process of the prototype vector of a vowel ($/v/$) category $\mathbf{r}_d^{/v/}$

learning. Therefore, we provide the robot with a linear transformation as a mapping function from the region of generable sounds by the caregiver to that 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 corresponding sound $\mathbf{h'}^{/v/}$ 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 the generable regions by the caregiver and the robot, respectively.

B. Updating the 'vowel' categories of the robot

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

Suppose that the robot utters $\mathbf{r}_d^{/v/}$ which is one of the current prototype vowels 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 $\mathbf{h}^{/v/}$. These processes are formalized as follows:

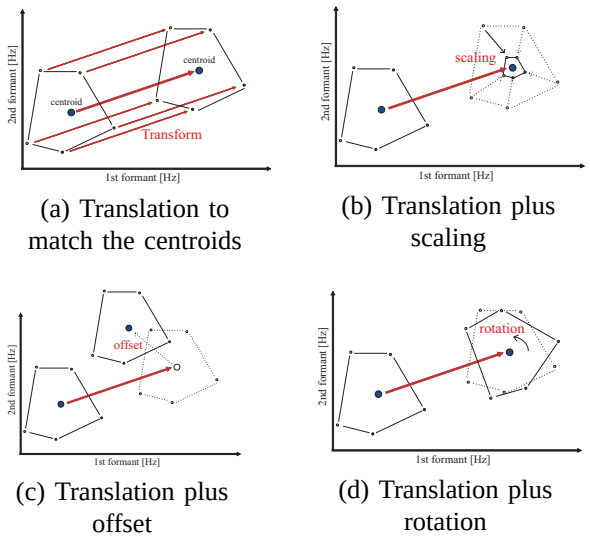
1) At the k-th step, the robot selects one of vowels $/v/$ and utters it with the current prototype category $\mathbf{r}_d^{/v/}(k)$.
2) The caregiver generates the sound $\mathbf{h}(k)$ to imitate the robot's utterance.
3) The difference vector $\Delta\mathbf{h} = \mathbf{h}^{/v/} - \mathbf{h}(k)$ is converted to the region of generable sounds by the robot with the mapping function $\mathbf{g}$. The converted difference vector is applied to modify the prototype category $\mathbf{r}_d^{/v/}(k)$, in other words,

$$\mathbf{r}_d^{/v/}(k+1) = \mathbf{r}_d^{/v/}(k) + \mathbf{g}(\Delta\mathbf{h}) \quad (2)$$

Figure 3 illustrates these processes schematically.
4) Again, the robot utters the voice with the new prototype category $\mathbf{r}_d^{/v/}(k+1)$.

IV. Experiment

In the experiments, we 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



(a) Translation to match the centroids

(b) Translation plus scaling

(c) Translation plus offset

(d) Translation plus rotation

Fig. 4. Four examples of mapping functions from the region of generable sounds by the caregiver to that by the robot

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 maternal imitation are more acceptable as Japanese vowels than those acquired from fixed desired formant vectors.

We used four types of rough estimation in the experiments, as shown in Fig. 4: (a) translation to match the centroids, (b) translation plus scaling, (c) translation plus offset, and (d) translation plus rotation.

First, we describe our vocal robot and the method by which it forms utterances. Next, the experimental procedures are explained, and the results of the imitated and acquired vowels with statistical analysis are then given.

A. Vocal robot

Vocalization is commonly regarded as the result of 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 as the source-filter theory of speech production [12] and has been implemented in previous studies [7], [6]. To model the process of vowel convergence in mother-infant interaction, we improved the vocal robot used in a previous study [7] such that the sound source was replaced by an air

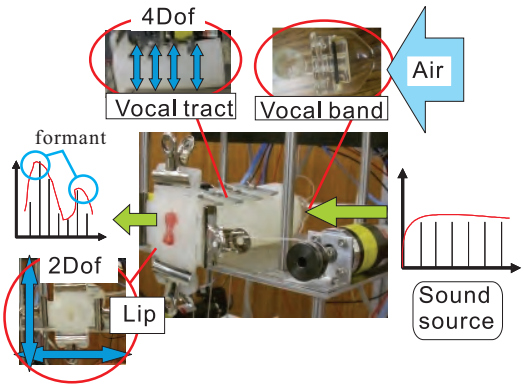


Fig. 5. Articulatory system of the vocal robot

TABLE I

MOTOR COMMANDS TO FORM LIP SHAPES THAT RESEMBLE THOSE OF A HUMAN 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

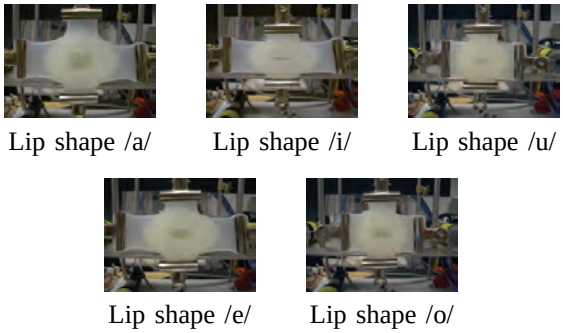


Fig. 6. Lip shapes of the robot

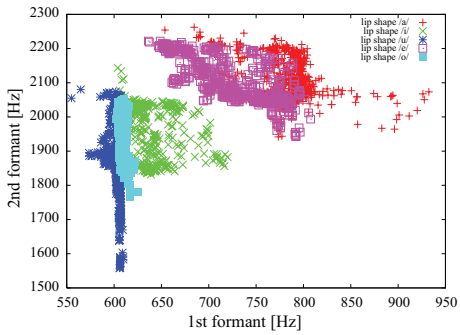


Fig. 7. Distribution of robot utterances in the formant space

compressor and an artificial vocal band. In addition, a lip was added at the front end of the vocal tract, and the length of the robot's vocal tract changes from 170 [mm] (average male 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, the sound wave is then spread out through the vocal tract and the lip, which is a silicon tube with a hollow end, thus resembling a human lip. To modulate the sound wave, the vocal tract and the lip were wired with four electric motors, respectively, by which they could be deformed. The host computer controls the motors through motor controllers (us-biMC01, iXsResearch 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 the lips by four motors, and four of which are used for deforming the vocal tract by another set of motors. First, we show the utterance capability of the robot. The motor commands that control the shape of the vocal tract are quantized into five levels, 0 (free, no deformation), 0.25, 0.5 (medium), 0.75, and 1.0 (maximum deformation), and the motor commands that control the lip shape are assigned to imitate the shape of human lips. Table I and Fig. 6 show the motor commands and lip shapes of the robot used to imitate the human lip shape. Figure 7 shows the formant distribution of the robot utterances in the formant space, where the horizontal and vertical axes indicate the first and the second formants, respectively.

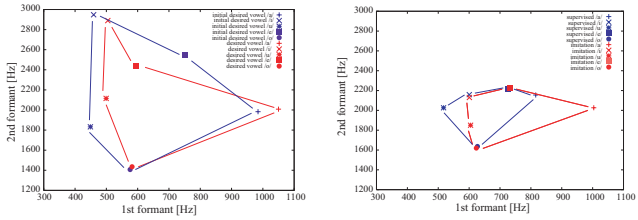
Using the data in Fig. 7 as the list of the pairs of the motor commands and the formant vectors, the robot can generate the desired sound. From the list of the pairs, the robot can find a number of candidate pairs for which the formant vectors are sufficiently close to the desired vectors. The robot then selects from the candidates a pair that has the closest motor command to the previous motor command.

B. Setup and procedure

The experiments are conducted under the condition that one subject (the same caregiver throughout all of the experiments) participated in the vowel acquisition process by two methods using four mapping functions each. That is, a total of eight experiments were conducted to verify the hypotheses. Note that each experiment is iterated five times for later statistical analysis. In the vowel acquisition process with the proposed method of maternal imitation, the caregiver tries to imitate the robot's utterances such that other person would judge the imitated voice as being the same as that of the robot. During the alternation of the uttering voice, the robot modifies the desired formant vectors using the caregiver's utterances as the information of the correspondence of both utterances. For 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 four mapping functions are as follows:

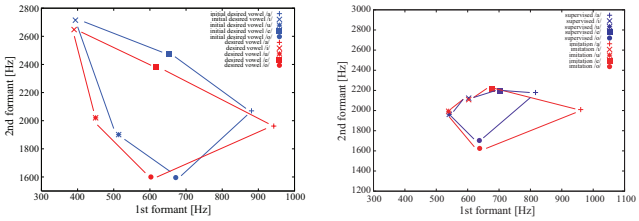
- translation: only translation by the difference between two centroids: α=1.0, **R**(0), **s**=(0, 0) (See Fig. 4 (a)).
- scaling: translation to match the centroids plus scaling: α=0.24 (that coincides with the region of the generable sounds of the robot), **R**(0), **s**=(0, 0) (See Fig. 4 (b)).
- offset: translation to match the centroids plus offset: α=1.0, **R**(0), **s**=(-100, 200) (See Fig. 4 (c)).
- rotation: translation to match the centroids plus rotation: α=1.0, **R**(30), **s**=(0, 0) (See Fig. 4 (d)).

The mapping function of 'translation' is regarded as one of simple and feasible translation to match vowels in different regions of formant space. The other mappings are example varieties which contain some noise in such a feasible



(a) Desired formant vectors (b) Acquired formant vectors

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



(a) Desired formant vectors (b) Acquired formant vectors

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

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

C. Results

First, we present the robot's vowels that were acquired through the 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 Fig. 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 denoted by blue symbols (+, * etc.) and red symbols, respectively. Hereafter, blue and red indicate data obtained by supervised learning and maternal imitation, respectively. In Fig. 8(b), the vowel categories as formant vectors acquired by both methods are indicated in the same colors as in Fig. 8(a). Figures 9(a) and 9(b) show similar graphs to those of Figs. 8(a) and 8(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 Fig. 8 and Fig. 9 imply that the robot succeeded in modifying its desired formant vectors.

We hypothesized that unconscious anchoring gradually leads the caregiver's utterance to his/her own vowels, and, in order to verify this tendency, the changes in the difference $\Delta\mathbf{h}$ in Fig. 3 (the distance indicating the error of the mapping) at the beginning and at the end of the learning are examined. This change is shown in Fig. 10. In the figure, the vertical axis indicates the size of $\Delta\mathbf{h}$, and the vertical bars indicate the average of first three instances of learning and the average of the last three instances of learning that were acquired through five experiments using each mapping function. In addition, the narrow bars indicate the standard deviation of instances. A T-test indicated a highly significant

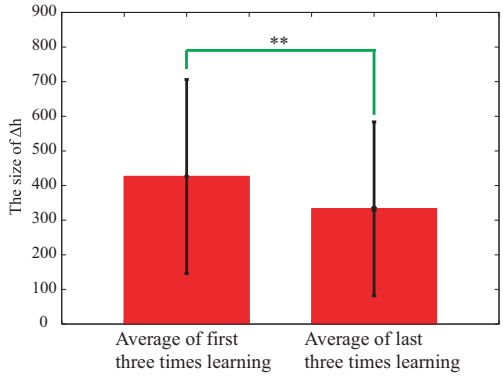


Fig. 10. Difference between the imitated voices by the caregiver and his/her corresponding usual vowels at the beginning and the end of the interaction

difference in the average size of $\Delta\mathbf{h}$ between the first three steps and 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 verify the first hypothesis.

Since it is difficult to show the more vowel-like method between two methods (the maternal imitation and the supervised learning) in the formant space, we used a subjective criterion to judge the more vowel-like method. We asked 15 subjects to compare the vowels acquired through maternal imitation with vowels acquired by the supervised learning. In this test, the robot continuously utters vowels acquired through maternal imitation or with the supervised learning with four different mapping functions in the normal order of Japanese vowels, that is /a/,/i/,/u/,/e/,/o/. Subjects were asked to compare the robot's vowels four times (two sets of the voices by the maternal imitation followed by that of the supervised learning, and vise versa) in terms of four mapping functions and to judge which vowels were more vowel-like in terms of being recognizable as Japanese vowels.

Figure 11 shows the results of a comparison of 15 subjects, where the percentage of subjects who reported the vowels acquired by maternal imitation to be better with each mapping function and the total percentage among all four mapping functions are denoted by red bars, and the percentage of subjects who reported the vowels acquired by supervised learning to be better are denoted by blue bars. We conducted tests to determine whether the percentage of subjects who positively answer on the utterances acquired by the maternal imitation is higher than the chance level (50%). From statistical tests, we found that the subjects tend to judge the utterances acquired by maternal imitation to be more vowel-like than those acquired by supervised learning for three mapping functions, namely translation ($p = 1.6 \times 10^{-3}$), offset ($p = 4.7 \times 10^{-5}$), and rotation ($p = 2.2 \times 10^{-2}$). Although there was no significant difference with the chance level in the case of a mapping function of scaling ($p = 2.8 \times 10^{-1}$), the test on the total percentage among four mapping functions indicated the tendency of the utterances acquired by maternal imitation to be more vowel-like ($p = 1.1 \times 10^{-2}$). Based on the comparison, we conclude that the robot was able to acquire better vowels

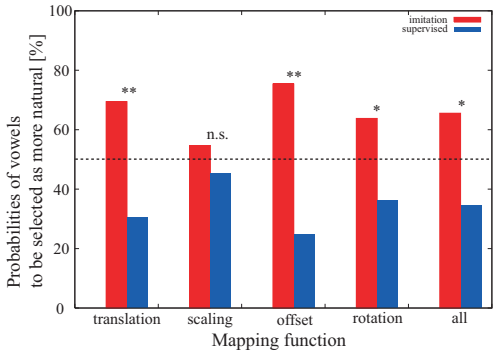


Fig. 11. Probabilities of vowels to be selected as more natural

by maternal imitation or, in the worst case, vowels that were recognizable by a human as being Japanese vowels, regardless of the mapping functions.

V. Discussion and Conclusion

In this paper, we proposed a vowel learning method with an imitative caregiver based on the idea of unconscious anchoring behavior of humans. In experiments, a vocal robot succeeded in acquiring more vowel-like sounds through iterative turn-taking with a caregiver who tried to imitate the robot's utterances. Even though the caregiver did not intend to instruct the robot, intending rather to simply imitate the robot's sound, the human tendency of unconscious anchoring worked to lead the robot's utterances to be more vowel-like. Since unconscious anchoring is expected to provide a new paradigm by which to provide a robot with human-like behaviors, in the future, we intend to quantitatively investigate the extent to which such a tendency can be expected and can be used for this purpose.

One of the most fundamental issues in imitation is to find the mapping function from the observation of the behaviors of others to one's own behaviors. In the present study, the mapping function is approximated by an affine function in the formant space. The parameters used in the experiments are limited, and we have not examined other parameters. Estimates of the affine approximation that are exceedingly incorrect do not seem work, but we suppose that the parameters might work, unless the vowel categories interfere with each other by the transformation for modification. If the vowel categories interfere with each other, then the desired formant vector jumped to the wrong category. How can we guarantee no interfere is one of our future issues.

In the present study, the parameters of the mapping function do not change during the maternal imitation process, although these parameters are not exactly correct. Therefore, the learning (modification) of these parameters simultaneously with modification of the desired vectors is a natural extension of the current work, and we conjecture that this is the process by which an infant, during the cooing process, learns the vowel categories based on the experience of listening to his/her mother's utterances. Supporting evidence is partially observed in developmental psychology [13], [14]. However, real infants are exposed not simply by single vowels, but rather by continuous voices with consonants, that is, words, phrase, and sentences. Furthermore, mothers do not simply respond by imitating their infants' utterances. In such an environment, the extension of the present research is very challenging.

Since "unconscious anchoring" is considered to be the general concept of human behavior in imitation, the framework of the present study is expected to be applied to other modalities, such as vision (and motion, that is, gesturing). To show this generality is another future issue. Since the multi-modal senses of the human are said to be interfere with each other (e.g., the McGurk effect [15]), another possibility of the extension would concern the hybrid effects of unconscious anchoring, not only on hearing, but also on sight. Although the effects of these modalities were not well separated in the preset study, investigations to separate these effects is an important issue, and hopefully the requirements of the body or the appearance of the robot will effectively utilize multi-modal unconscious anchoring.

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*, Vol. 37, pp. 185–193, 2001.

[3] B. de Boer. Self-organization in vowel systems. *Journal of Phonetics*, Vol. 28, pp. 441–465, 2000.

[4] P.-Y. Oudeyer. Phonemic coding might result from sensory-motor coupling dynamics. In *Proceedings of the 7th international conference on simulation of adaptive behavior (SAB02)*, pp. 406–416, 2002.

[5] 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*, pp. 2526–2531, Augst 2005.

[6] 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*, pp. 3858–3863, 2002.

[7] Yuichiro Yoshikawa, Minoru Asada, 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, Dec 2003.

[8] 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, p. 670, 1996.

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

[10] Katsushi Miura, Minoru Asada, Koh Hosoda, and Yuichiro Yoshikawa. Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction. In *The 5th International Conference on Development and Learning (ICDL'06)*, 2006.

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

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

[13] Anthony J. DeCasper and Melanie J. Spence. Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, Vol. 9, pp. 133–150, 1986.

[14] Patricia K. Kuhl. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, Vol. 66, pp. 1668–1679, 1979.

[15] Harry McGurk and Jrohn MacDonald. Hearing lips and seeing voices. *Nature*, Vol. 72, pp. 746–748, 1976.