Vowel Acquisition based on Visual and Auditory Mutual Imitation in Mother-Infant Interaction

Katsushi Miura, Minoru Asada, Koh Hosoda, and Yuichiro Yoshikawa
JST ERATO Asada Synergistic Intelligence Project (www.jeap.org)
Graduate School of Engineering, Osaka University
{*miura,asada,hosoda,yoshikawa*}*@er.ams.eng.osaka-u.ac.jp*

***Abstract*— A pioneering constructivist approach to building a robot that reproduces a developmental process of infants' vowel acquisition has been conducted by Yoshikawa et al. [1] inspired by the observation in infant study. They have constructed a mother-infant interaction model with robot learning capability and parrot-like teaching by caregiver. However, the robot has not listened his/her own voice, therefore it could not actively explore more natural vowels similar to the caregiver.**

The study presented in this paper extends the previous work in the following manners seeking for more natural interaction. First, a lip is added to the robot to imitate the lip shape of the caregiver in order to accelerate the learning process by constraining the initial exploration area in the formant space. Second, the pentagon the caregiver's vowels construct in the formant space is utilized as the desired vowels for the robot. Third, mutual imitation between the robot and the caregiver is introduced in order to obtain more natural vowels hypothesizing that the caregiver imitates the robot voice but unconsciously the imitated voice is close to one of his/her own vowels. Through this process, the desired positions specified at the second step is gradually shifted, expecting to more natural ones. The experimental results are shown and the future issues are discussed.

***Index Terms*— Vowel Imitation, Formant Space, Lip Shape, Maternal Imitation**

I. Introduction

Vocal communication with humans is one of the most formidable challenges in humanoid robotics, and how human infants acquire this function is also one of the mysteries of human cognitive development. A constructive approach to understanding this process by building a robot that can reproduce the vocalization seems a promising way as an approach from cognitive developmental robotics [2].

An infant robot is supposed to have a capability to acquire phonemes without any knowledge about the relations between phonemes and its sensorimotor system as human infants are supposed. Thus, the robot must obtain information for learning them through interactions with its environment, namely its caregiver. Previous studies showed that a population of computer simulated agents with a vocal tract and cochlea can acquire shared vowels by self-organization through interactions with each other [3], [4]. Although they did not assume *a priori* knowledge about vowels, there was an assumption that the agents can reproduce sounds similar to those of other agents so that "imitation game [3]" or "magnet effect [4]" leads to share vowels in population. However, we should take infant immaturity into account for modelling the vowel acquisition process since infants cannot reproduce the caregiver's utterances as they are.

Yoshikawa et al. [1] proposed a mother-infant interaction model for infant vowel acquisition inspired by the study in infant development that maternal imitation effectively reinforces infant vocalization [5] and that its speech-like cooing tends to lead utterances of its mother [6]. They hypothesized that imitation by the caregiver, which is repetition of infant's vocalization with adult phonemes, plays an important role in phoneme acquisition through interactions, and implemented the model with learning capability and parrot-like teaching by caregiver. A vocal robot they built obtained four of five Japanese vowels through interactions with caregiver based on the hypothesis. However, the robot has not listened to his/her own voice, therefore nor actively explored more natural vowels similar to the caregivers'.

In this paper, we extends the previous work in the following manners seeking for more natural interaction. First, a lip is added to the robot to imitate the lip shape of the caregiver in order to accelerate the learning process by constraining the initial exploration area in the formant space which is a well-known sound feature space to distinguish vowels [7]. From the lip shape imitation, the initial locations of vowels are obtained. Second, the pentagon the caregiver's vowels construct in the formant space is utilized as the desired vowels for the robot by shifting this pentagon to the centroid of the initial vowel locations of the robot. Third, mutual imitation between the robot and the caregiver is introduced in order to obtain more natural vowels hypothesizing that the caregiver imitates the robot voice but unconsciously the imitated voice is close to one of his/her own vowels. Through this process, the desired positions specified at the second step are gradually shifted, expecting to more natural ones. The experimental results are shown and the future issues are discussed.

II. Vowel imitation based on mother-infant interatction model

An overview of the whole system is shown in Fig. 1 where a vocal robot with articulation and auditory functions interacts with a caregiver who shows lip shape to the robot, and both imitate their voices each other. We slightly modified the vocal robot in [1] by adding the lip structure that has two degrees of freedom corresponding to opening/closing in horizontally and

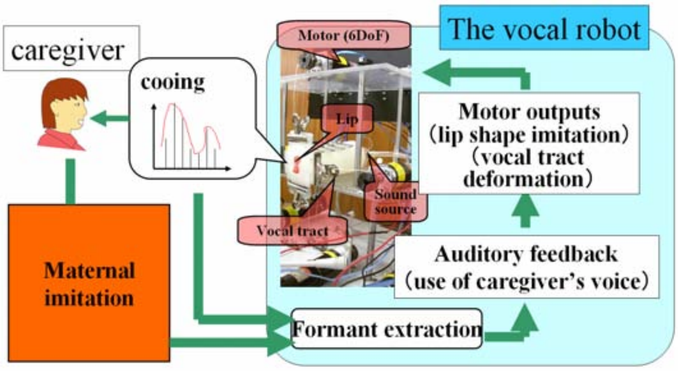


Fig. 1. An overview of the whole system

vertically. The shape change of the vocal tract is controlled by four motors instead of five ones in the previous work.

The caregiver has two roles: the first one is to show the lip shape to the robot so that it can imitate the lip shape since there are some evidences on the relationship between the lip shape and the vowel utterance. Carre [8] shows the formant changes owing to opening/closing of the lip by his simulation. Patterson and Werker [9] found that the infants have knowledge on the relationship between the lip shape and phoneme. Then, we assume that the robot can imitate the lip shape corresponding to each vowel from the caregiver to initialize its vowels in the formant space for further exploration.

The second one is maternal imitation, that is, imitation of the robot voices. Since the robot also imitates the caregiver's imitation, mutual imitation is expected to lead the robot voices to more natural one. The implicit assumption behind this process is that the caregiver tries to imitate the robot voices, but unconsciously his or her voices are close to his or her own vowels owing to the embodiment (sensorimotor constraint). As a result, the robot voices gradually change to more natural ones.

In the followiongs, first how the lip shape imitation constrain the vowel exploration is shown, next the use of the shape of the caregivers' vowels in the formant space is explained, and then the learning method with maternal imitation is given.

III. Lip shape imitation for vowel acquisition

Visual imitation for vowel acquisition is realized by the lip shape imitation since the relationship between the lip shape and the utterance is found [8], [9]. Here, we show how this imitation constratin the vowel exploration in the formant space, and the similarity in relative placement of vowels between human and the robot in the formant space.

A. The relationship between the lip shape and formants

The vocal robot has six degrees of freedom, two of which are for lip opening/closing. First, we examine the utterance capability of the robot. The motor command is normalized into three levels 0 (free, no deformation), 0.5 (middle), and 1.0 (the maximum deformation), therefore we have 729 (3^6) utterances. Among them, we picked up 711 utterances after removing unstable onses. Fig.2 shows their plots in the formant space where the horizontal and vertical axes indicate the first and second formants, respectively.

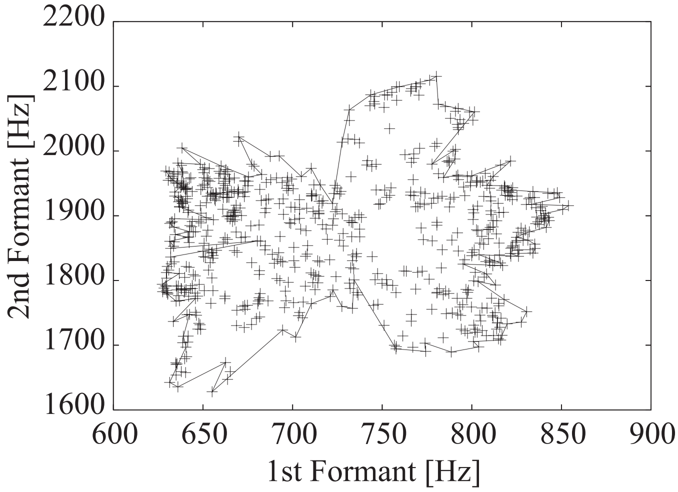


Fig. 2. Formant distribution of robot's voice

TABLE I

Relation between robot's lip shape and motor outputs

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

Next, the effect of the lip shape is examined, that is, two of six DoFs corresponding lip opening/closing is fixed to imitate the human lip shape, and other four DoFs are combined (81 utterances, 3^4). Table I shows the motor commands that imitate the human lip shape, and Fig.3 shows the formant distribution of the robot utterances when the lip shape is imitated. The top row indicates the formant distributions of utterances when the lip shape is fixed to imitate that of human's vowel utternaces /a/, /i/, /u/, /e/, and /o/, respectively. The bottom row shows the lip shapes of human and the robot, respectively.

The number of utterances constrained by lip shape is 394 (11 unstable utterances are removed) from 711 without the constraint, comparing Figs. **??**. This tells that by lip shape imitation, the number of candidates for vowel utterances can be reduced to 55%.

B. The formant placement of human's vowels and the robot ones

The placement of the human's vowels in the first and second formant space is shown in Fig.4, where the relative position among five vowels /a/, /i/, /u/, /e/, and /o/ constructs a pentagon. This relative positions is similar to that of the robot vowels (see Fig.3). For example, the first formant of /a/ is high, and its second formant is around the center of the all vowels.

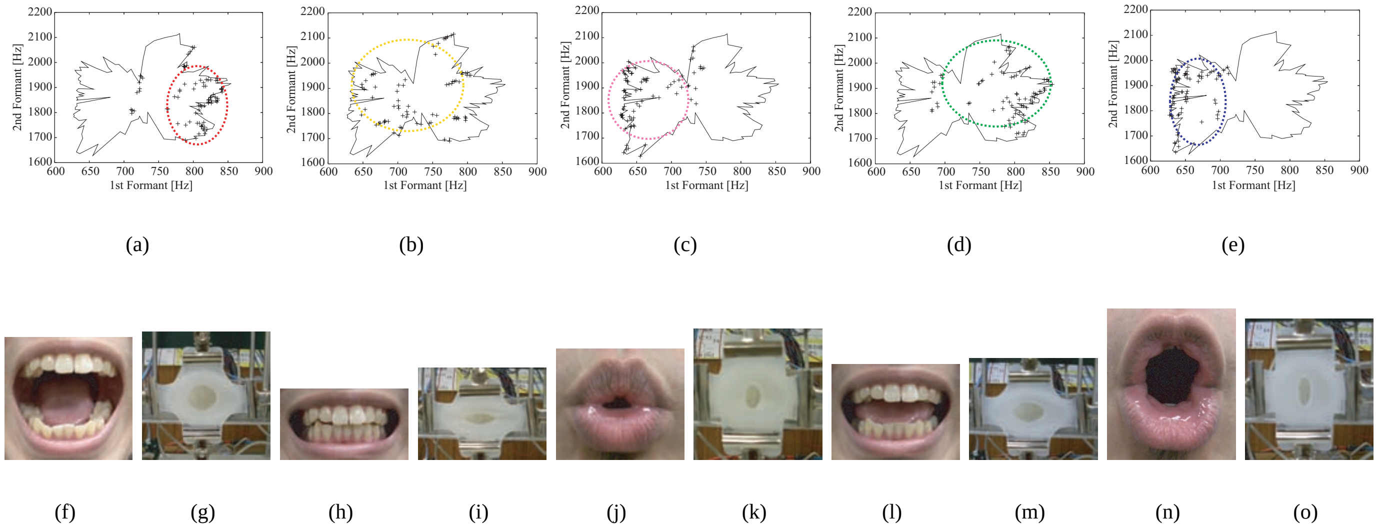


Fig. 3. Formant distribution of the utterances (top: /a/, /i/, /u/, /e/, /o/) from lip shape imitation (bottom)

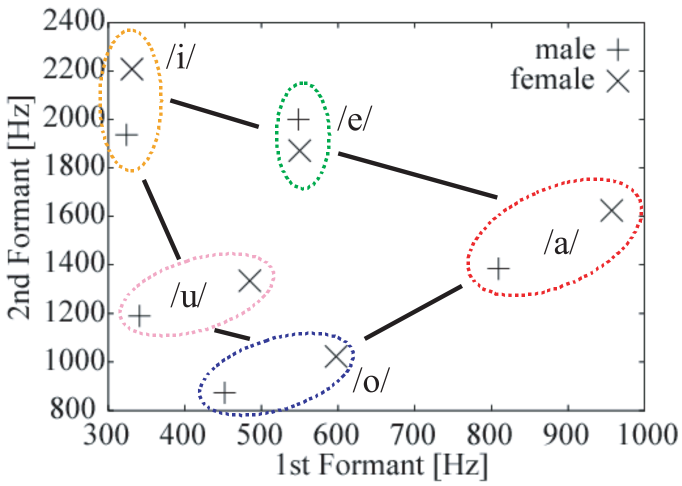


Fig. 4. Formant distribution of human's vowel

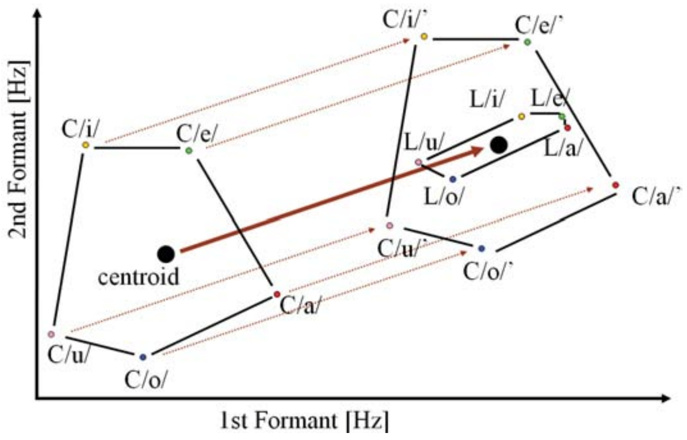


Fig. 5. Transformation of human's vowels to the robot utterance area

This implies that the human's vowel positions might be useful for the robot to regard them as goals to utter more natural vowels, but the dynamic range and also the centroid of both distributions are differnet. Then, we move the human's vowel positions (C/a/-/o/)to the robot vowels area (L/a/-/o/) so that the both centroids coincident with each other as shown in Fig.5. The shifted human vowels (C/a/'-/o/')can be the desired positions for the robot to achieve to obtain more natural vowels.

C. Learning to acquire more natural vowels through mutual imitation

From the above discussion, the exploration process to find more natural vowel utterance seems easy since the robot can follow the trajectory between its initial position (L/j/, j=a, i, u, e, or o) to the desired vowel position (C/j/', j=a, i, u, e, or o) under the condition that the lip shape is fixed. However, there is no guarantee that the desired vowels can be heard as natural ones. Then, we introduce mutual imitation to shift the desired position so that the robot can acquire more natural ones by hypothesizing that the caregiver's imitation is biased by his/her own sensorimotor constraint, that is, embodiment, and therefore unconsciously implies the direction of natural utterance in the formant space.

Fig.6 explains how mutual imitation can improve the robot utterance. Let D/j/[k], M/j/[k], C/j/[k]$_{imi}$ (j=a, i, u, e, or o) be the desired position, the current position and the utterance that caregiver imitates at the time=k, respectively. Initially, D/j/[0] = C/j/' and M/j/[0] = L/j/, and C/j/ is the caregiver's original vowel.

1) The robot finds the the current position M/j/[k] on the

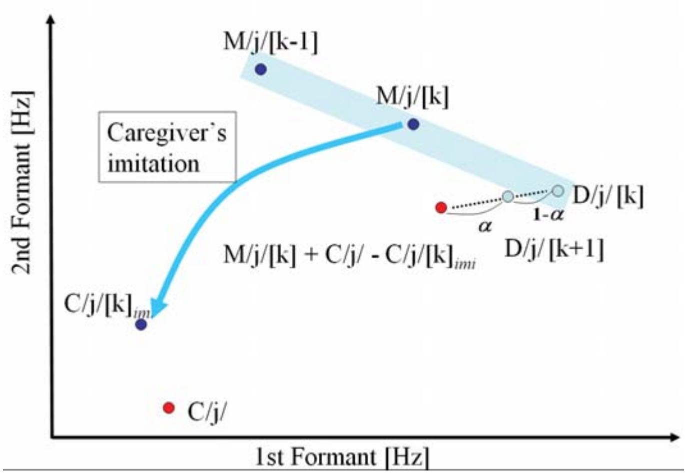


Fig. 6. Mutual imitation process on the formant space

trajectory between M/j/[k-1] and D/j/[k], and utters this position to the caregiver.

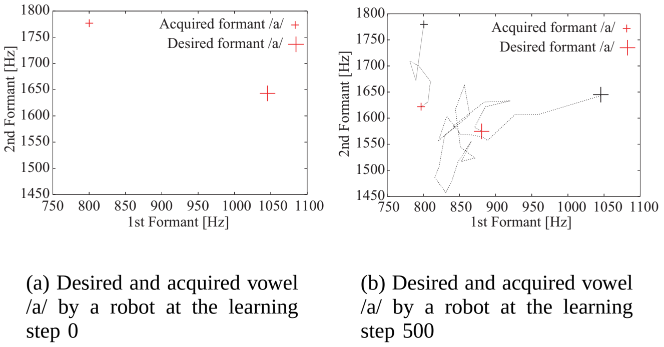
2) The caregiver imitates the robot utterance M/j/[k], but the imitated utterance C/j/[k]$_{imi}$ is close to his/her own vowel C/j/ owing to the sensorimotor constraint.
3) A difference vector from C/j/[k]$_{imi}$ to C/j/ is supposed to indicate the direction to more natural vowel utternace.
4) Then, the desired position of the utterance is set at the position M/j/[k] + C/j/ - C/j/[k]$_{imi}$ from the current position.
5) To avoid a big change, a new desired position D/j/[k+1] is set as follows: D/j/[k+1] = α D/j/[k] + (1-α) (M/j/[k] + C/j/ - C/j/[k]$_{imi}$).
6) Repeat the above process until M/j/ does not change (often, due to the limit of the utterance capability).

IV. Experimental Results

We apply the learning algorithm to the real robot and a caregiver with $\alpha = .7$. Here, we stop the learning at the 500th step since almost no changes happened in the robot utterance. Figs.7 (a) and (b) show the changes of the desired and acquired positions in the formant space. In (a), the initial positions, that is, M/a/[0] (small +) and D/a/[0] = C/a/' (large +) are indicated, and in (b), the final positions M/a/[500] (small +) and D/a/[500] (large +) are indicated.

Fig.8 shows the results of all vowels where thin folded lines of D/j/ indicate how they are modified according to the mutual imitation. The initial D/j/s are shown as nodes of a pentagon connected by dotted black lines and the final ones as nodes of a pentagon connected by dotted yellow-green lines. The acquired vowels are shown as nodes of a pentagon connected by dotted sky-blue lines. The initial nodes are black and the final ones colored. The human vowels are shown as larger marks such as +, *, and so on, while the robot ones as smaller ones.

Since it is difficult to show how natural the robot utterances are, we prepared the robot utterances without mutual imitation, that is, the robot acquired the vowels supposing the initial desired positions D/j/[0] untl the end of learning, and asked about 40 naive people (no knowledge on vowel learning). As a result, 70% agreed that the utterances with mutual imitation are more natural than that without mutual imitation.



(a) Desired and acquired vowel /a/ by a robot at the learning step 0

(b) Desired and acquired vowel /a/ by a robot at the learning step 500

Fig. 7. The learning process in the case of the vowel /a/

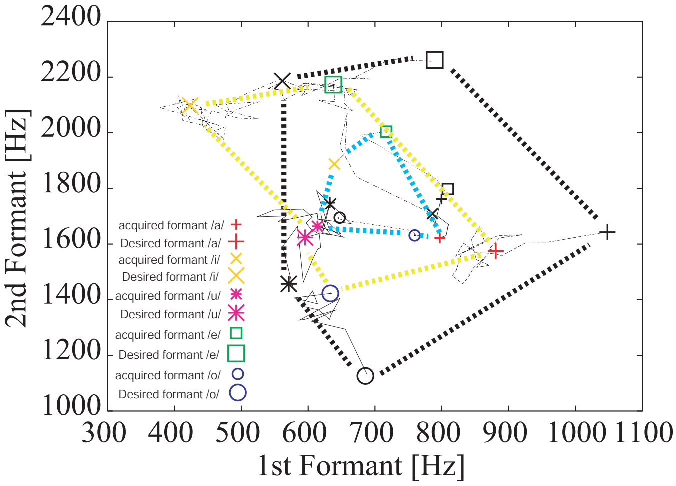


Fig. 8. Desired and acquired vowels by a robot at the learning step 500

V. Discussion

Currently, we emulated the lip shape imitation by hand. In a real situation, the robot should capture the caregiver's face through its TV camera and extract the lip shape information. Then, it should send the motor commands to imitate the observed lip shape. To do that, the robot should know the correspondence between the body parts such as eyes, nose, and lip, and also how to control its body part, in this case lip motion. Neonatal imitation tells us such a capability [10] though it has been controversy. However, its mechanism has not been revealed yet. Development of boy representation and motor control is one of the key issues in cognitive developmental robotics, and we will attack this problem in near future.

Another issue is how the caregiver behaviors affect the learning results. In this paper, the number of the caregiver is just one, but we have not recorded how he responded (imitated) to the robot. The analysis of such interaction data would be helpful to understand the developmental process of vowel acquisition, and also to build the design policy of the learning robots.

In our study, we adopted the following hypotheses that the lip shape information helps the infant's vowel acquisition and that the caregiver's imitation tends to similar to his/her own vowels. Based on these, we build the interaction model and showed the experimental results. The verification for these hypotheses is not complete. Collaboration with infant study and developmental psychology seems necessary not only to verify them but also to refine or modify the model.

REFERENCES

[1] 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.

[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] B. de Boer. Self-organization in vowel systems. *Journal of Phonetics*, 28:441–465, 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)*, pages 406–416, 2002.

[5] 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.

[6] 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.

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

[8] Carre R. Prediction of vowel systems using a deductive approach. In *In Proceedings of the International Conference on Spoken Language Processing 96*, pages 434–437, 1996.

[9] M. L. Patterson and J. F. Werker. Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6(2):191–196, 2003.

[10] Andrew N. Meltzoff and M. Keith Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, pages 179–192, 1997.