Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching

Y. Yoshikawa*, J. Koga*, M. Asada*†, and K. Hosoda*†

* *Dept. of Adaptive Machine Systems,* † *Handai Frontier Research Center,*
Graduate School of Engineering, Osaka University

Abstract

Without any explicit knowledge, human infants acquire the phonemes of adults who have different articulation parameters. By building a robots that reproduce this developmental process, we expect to model as yet unknown cognitive developmental processes and to discover fundamental design principles for machines capable of vocal communication. This paper proposes a model for acquring Japanese vowels based on observations of imitation between a mother and an infant. We tested the validity of the proposed model by examining whether a real robot can acquire vowels through interactions with a caregiver.

1 Introduction

Vocalization is one of the most promising methods of human-robot communication from the viewpoint of efficiency of information transfer. For an agent to communicate by vocalization, it should share common phonemes, which are abstract units of the phonetic system of a language, with its interlocutor. Human infants acquire phonemes of their mother tongue and finally their mother tongue itself through interaction with their caregivers including interaction involving babbling. In this study, we aim to build a robot that learns to acquire a capability of communication by vocalization with human infants based on a constructivist approach, that is, having it develop in a way that is similar to human development. As Asada et al. have suggested in their discussion of cognitive developmental robotics [1], building this kind of a robot may help us model the human developmental process of acquiring phonemes.

Nishikawa et al. [2] built an anthropomorphic robot that can produce Japanese phonemes including consonant sounds. However, its designers manually tuned phoneme articulation because they have not addressed the acquisition problem itself. Higashimoto et al. [3] built a system that learns to vocalize human vowels. It consists of a vibrator, which generates a source sound, and a deformable silicone rubber tube that works as a filter to change the spectrum envelop of the source sound. It can vocalize human-like vowels by an inverse model of articulation parameters with respect to a spectrum envelop of sound. The generated sound is modified to more closely resemble reference vowel uttered by the experimenter. In other words, it imitates according to the raw soundwave similarity of its vocalizations to human vocalizations. Vocalization based on playback systems, for example [4], could also be regarded as imitation based on the similarity of raw soundwaves. However, human infants can imitate adult phonemes in spite of the fact they cannot vocalize the same sounds from the viewpoint of raw soundwave similarity since their vocal mechanism is immature. Therefore, we argue that a robot should acquire phonemes by a different method of imitation since raw soundwave imitation cannot explanin how human beings acquire phonemes.

We assume that a robot can acquire phonemes without any knowledge about the relations between phonemes and its sensorimotor system. Thus, it must obtain information for learning them through interactions with its environment, namely its caregiver. Therefore, we have two main design issues: 1) what kind of mechanism for interaction should the robot possess and 2) how can the caregiver fasilitate the learning of phonemes. We observe that maternal imitation effectively reinforces infant vocalization [5], and therefore we hypothesize that imitation by the caregiver, that is, parrot-like imitation of the infant's vocalizations, plays an important role in phoneme acquisition. The purpose of this study is to build a robot that acquires phonemes through random vocal articulations and interactions with a caregiver who parrots the robot's vocalizations.

In this paper, we address the problem of acquiring five Japanese vowels by using a robot that can articulate its sounds by deforming its silicon vocal tract and has a similar vocal apparatus to the one used in Higashimoto et al. [3]. We assume that the robot has an innate capability of extracting formants, which are well-known sound features that are effective for distinguishing vowels. The learning mechanism consists of two layers and connections between them. One is the auditory layer for clustering formants of the caregiver by self-organization [6]. The other is the articulation layer, which clusters its own articulation parameters. The connections between them are updated by means of simple Hebbian learning through well designed interactions; therefore, the robot can acquire the caregiver's vowels even though the infant's articulation parameters are different. To simplify connections, we modified the learning rule by considering the facility of articulation.

The rest of this paper is organized as follows: First, we explain how to design interactions to enable the learning of vowels. Then we describe how learning works using the proposed method. After showing the configuration of the experimental robot and reporting a preliminary experiment to verify its acoustic properties, we examine whether the proposed method works.

2 The design of interaction

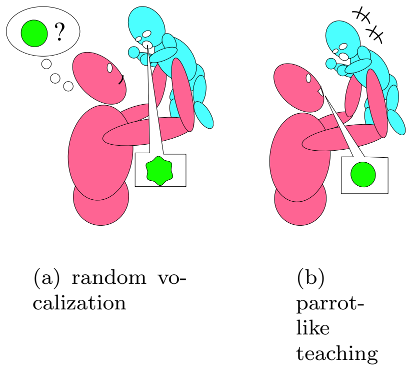
For a robot to learn to vocalize vowels without explicit knowledge about the relation between the vowels and its sensorimotor system, it should interact with a caregiver. Our approach is to create a situation in which robot and caregiver interact in a way that promotes the robot's learning.

Based on a study of mother-infant interaction [5], we conjecture that maternal imitation of an infant's vocalizations plays an key role in vowel acquisition. To build a robot that reproduces the observed interaction (see Fig. 1), we embed in the robot a random vocalization mechanism, which produces random infant cooing behavior. At the same time, the caregiver parrots the robot's vowel articulations, that is, the caregiver determines whether the robot has vocalized a vowel and then repeats it.

By designing the experimental situation in this way, the robot gains the invariant pairs of its articulation of vocalization and the corresponding vowel so that acquiring vowels can succeed with a simple learning rule despite the difference in the articulation parameters.

3 Learning mechanism

The robot's learning mechanism consists of two layers and connections between them (see Fig. 2). After describing the processing in these two layers, we give two learning rules for the connections between them.



(a) random vocalization

(b) parrot-like teaching

Figure 1: *Mother-infant interaction*

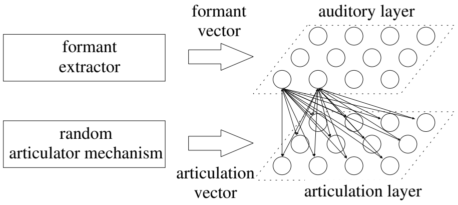


Figure 2: *Interconnected auditory and articulation layers constitute the learning mechanism.*

3.1 Auditory layer

Frequency peaks in the soundwaves of vowels are effective for distinguishing vowels; the peaks are called formants. The auditory layer receives *formant vectors* from the formant extractor. Each vector consists of the lowest four peaks of the caregivers' utterances; the auditory layer clusters them in a self-organizing manner, using a Kohonen map [6].

Let $\boldsymbol{f}_i = [f_{i1}, \cdots, f_{i4}] \in \Re^4, i = 1, \cdots, N_f$ be the code vector of the i-th unit in the auditory layer that consists of N_f units and $\boldsymbol{r}_i^f \in \Re^2$ be the topology vector of the same unit. When the auditory layer receives a formant vector $\boldsymbol{f} \in \Re^4$, units with closer code vectors become more active, and the most active unit suppresses the other units. Finally, an activation a_i^f of the i-th unit is calculated by

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

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

In the Kohonen map algorithm, code vectors of some units near the winner are modified to be closer 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 the time dependent scalar learning rate and $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is a monotonically decreasing function with respect to the distance between vector $\boldsymbol{x}$ and $\boldsymbol{y}$ that 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 that determines by how much nearby units learn according to a distance metric. At the start of learning, $\sigma(t)$ is set to such a high value that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is high in a wide region of the auditory layer and gradually decreases so that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ of nearby units only have high values at the end of learning. This learning rule clusters code vectors close to frequently observed input vectors in the auditory layer — in this case, frequently heard vowels.

3.2 Articulation layer

The articulation layer receives *articulation vectors* from the random articulation mechanism, which consists of five motor commands that deform the silicon vocal tract. The articulation layer clusters the vectors by the method as one used in 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 that consists of N_m units and $\boldsymbol{r}_i^m \in \Re^2$ be its i-th topology vector. When the articulation layer receives an articulation vector $\boldsymbol{m} \in \Re^5$, an activation a_i^m of i-th unit is calculated according to the eq. (1). Code vectors are updated acording to eq. (2). The most active unit is called the winner and labeled win^m.

3.3 Learning connections

The connections between the auditory layer and the articulation layer are updated based on the Hebbian learning rule that is used in artificial neural networks. Based on this learning model, connections between simultaneously active neurons in the auditory and articulatoin layers are strengthened while others are weakened. Let w_{ij} be a connection weight between the i-th unit in the auditory layer and j-th unit in the articulation layer. The learning rule 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 the 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 $a_i^f a_j^m$ [7]. We use the discrete version of the updating rule (eq. (4)) such that,

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

where t denotes the time interval.

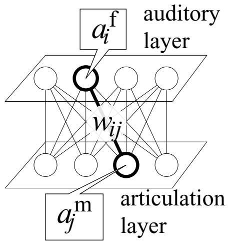


Figure 3: *Connections between the auditory and articulation layers.*

Employing the initially random articulation mechanism in parrot-like teaching causes invariant pairs of units to activate in both layers simultaneously. Therefore, through the learning process, articulations are matched with corresponding vowels as a connection between both layers.

However, interactions may connect multiple articulation units with a corresponding vowel since the caregiver will interpret some vocalizations caused by different articulations as the same vowel. To match a heard vowel with a unique articulation in order to vocalize it, we modify the learning rule so that it is responsive to the ease with which a vowel can be articulated — that is, more facile articulations are increased, while more difficult articulations are decreased. Therefore, the learning rule for the connections is slightly modified:

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

where $\eta(\boldsymbol{m})$ is an articulation facility function 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 the k-th element of the vector $\boldsymbol{x}$.

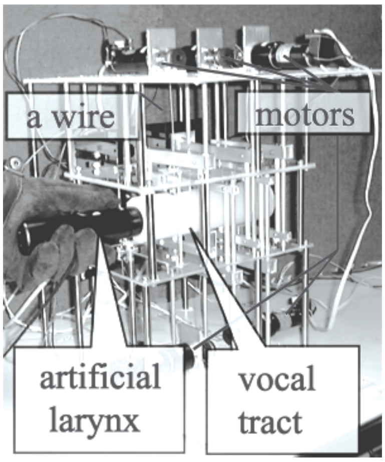


Figure 4: *The appearance of the experimental robot.*

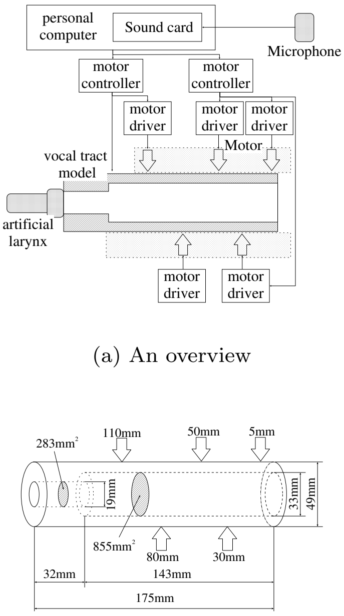
4 A robotic test bed

Vocalization is commonly considered to 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 as the "source-filter theory of speech production" [8]. We implement the source-filter theory by using a vibrator as a sound source and silicon rubber tube as a vocal tract that five electric motors deform. This implementation is similar to that of Higashimoto et al. [3] except that Higashimoto et al. use an artificial vocal cord while we use a membrane that a viabrator oscilates at a fundamental frequency.

Figure 4 and 5 depict the robot hardware. Five electric motors are bound to their respective attachments with piano wire. The motors pull at the attachments to deform the silicon vocal tract. They are controlled by motor controller (usbMC01, iXs Research Corp.) according to control commands from the host computer. An artificial larynx (Myvoice, Secom Medical System Co. Ltd.) is used as a sound source that generates a Rosenberg wave. The host computer receives signals from a microphone and calculates their formants.

4.1 Preliminary experiment

We conducted 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 the first, second, and third formant for vocalizations, which are interpreted as Japanese vowels by the experimenter. Figure 6 shows distributions of the calculated formants of the robot's vocalization. Formants of the experimenter are also shown for comparison. Table 1 shows averages for the robot and human vowels. The value in brackets is that of the experimenter.



(a) An overview

(b) Pulling points

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

We confirmed that the robot can vocalize four Japanese vowels but not /o/. Its distribution of formants tends to be higher than that of the human experimenter, but they are clustered in the space of formants. It means that formants are available for recognizing the vowels of the robot in addition to those of human beings.

Apparently, the robot cannot vocalize /o/ because its vocal tract does not have enough degrees of freedom. Therefore, the vowels in the following experiments exclude /o/.

5 Experiment

We conducted an experiment with the robot to test whether the proposed method works. After the robot vocalizes with the initially random articulation mechanism, the human caregiver determines whether the robot's vowel corresponds to the Japanese vowel and utters the corresponding vowel. The robot calculates formants of the caregiver's vocalization and updates the code vectors and connections according to the proposed method. To accelerate learning, the robot at first collects data from the interaction and then the learning mechanism is repeatedly trained with that data. Each element of a code vectors in the

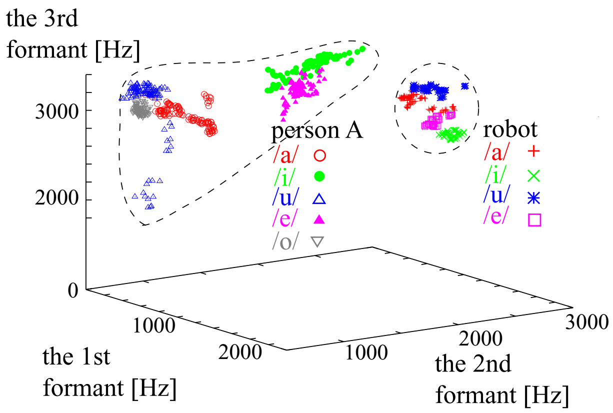


Figure 6: *Formant distributions of the robot and the caregiver(person A).*

Table 1: *Averages and standard deviation of formants of the robot and the caregiver (inside brackets).*

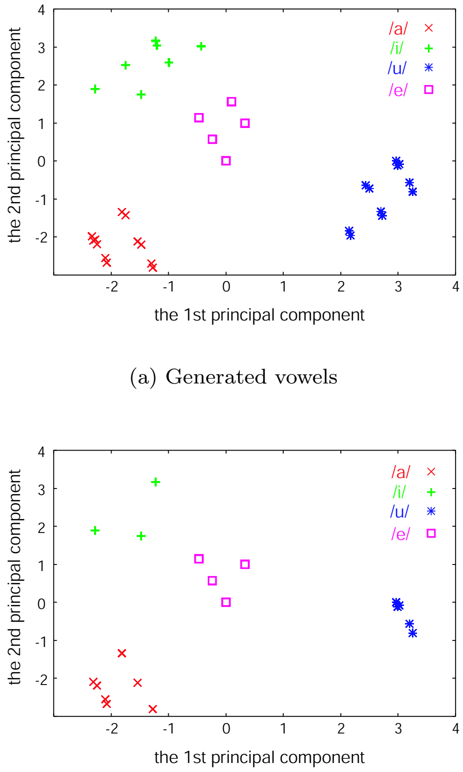
	1st [kHz]	2nd [kHz]	3rd [kHz]
/*a*/	1.33 ± 0.10 (0.40 ±0.05)	1.87 ± 0.15 (0.92 ±0.04)	2.64 ± 0.13 (2.93 ±0.08)
/*i*/	1.23 ±0.12 (0.23±0.02)	2.02 ±0.05 (1.77±0.28)	2.83±0.06 (3.11±0.09)
/*u*/	0.90 ±0.70 (0.25±0.03)	2.12±0.54 (0.96±0.10)	3.21 ±0.43 (2.78±0.21)
/*e*/	1.46±0.05 (0.31±0.11)	1.79±0.16 (1.24±0.38)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

articulation layer is quantized into five levels; these elements are the motor comands of the random articulation mechanism.

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 figure (a) and (b) are distributions of the articulation vectors compressed onto a two dimensional plane by the ordinal method of principal component analysis. Fig. 7(a) shows distributions of the articulation vectors of vocalizations that can be interpreted as Japanese vowels by the caregiver. When a caregiver utters a vowel, this activates units in the auditory layer, and this activation is propagated to the articulation layer via the stronger connections. Fig. 7(b) shows which units in the articulation layer are most strongly activated by the caregiver's vowels.

We can see that connected articulation vectors are parts of the region since the caregiver can interpret them as vowels corresponding to his or her utterances. Vocalizations that are generated by the articulation vector that is calculated by learned map are clear enough to be interpreted as vowels. Therefore, it is confirmed that the proposed method succeeded in learning Japanese vowels.



(a) Generated vowels

(b) Heard vowels

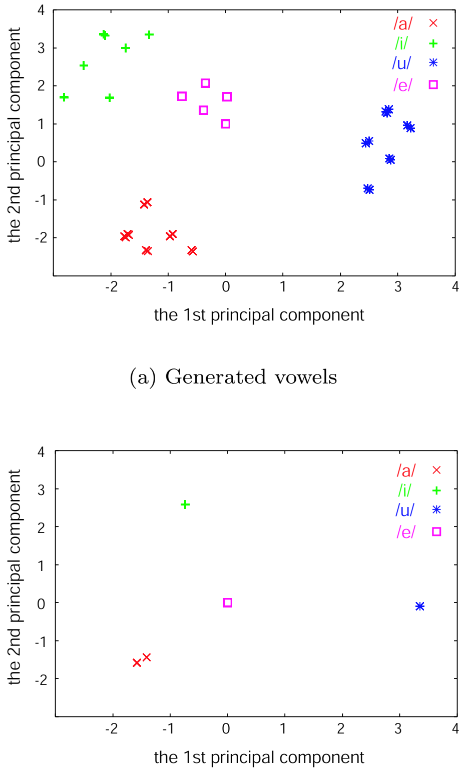
Figure 7: *Learning result without the facile criterion: articulation vector distributions.*

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 fewer articulation vectors are selected with respect to the caregiver's utterances than in Fig. 7(b). Therefore, we confirmed that the facility criterion decreased the number of units in the articulation layer that are generally activated by the auditory layer. The remaining articulation vectors are easy to articulate.

6 Discussion and Conclusion

In this paper, we propose a constructivist model of vowel acquisition between agents with different articulation parameters that does not depend on explicit knowledge. To acquire vowels, the caregiver-robot interaction depends on embedding a random articulation mechanism and adopting parrot-like teaching



(a) Generated vowels

(b) Heard vowels

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

based on studies of mother-infant interactions. An experimental robot is used to verify the proposed model.

When an agent tries to imitate the behavior of an agent with a different body structure, it needs to abstract observed behavior to some extent since it cannot duplicate it as it is. However, abstraction brings arbitrariness into the imitation process — even if the agent acquires pairs describing its own behavior and that of the caregiver. We proposed a method to cope with this arbitrariness by introducing a subjective criterion: the facility with which a vocalization can be articulated. As the facility criterion reduces arbitrariness of matched articulations for observed vowels in our second experiment, this kind of criterion could play an important role in imitation, understanding the behavior of others, and communicative processes between agents that have different bodies. Perhaps it is needless to say that any mechanism of communication must be able to account for the egocentric viewpoint of the agent.

Masataka [9] speculates that an infant can adjust the frequency trajectory of its vocalizations in reaction to its mother's responses from observation of contingent and non-contingent infant-mother interactions [9]. Although we need to observe the adaptations occuring in these interactions further, an infant's reactions seem to play a vital role in the process of acquiring phonemes and in the emergence of communication. However, our model cannot explain how infant's reactions play such roles since we apply a random articulation mechanism. Our future work will try to model how the interaction develops.

Acknowledgement

The Advanced and Innovational Research program in Life Sciences of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government and a Research Fellowship for Young Scientists from Japan Society for the Promotion of Science supported this research. We would like to thank Mr. Tsukinoki for his patient support in building our experimental robot.

References

[1] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous System*, Vol. 37, pp. 185–193, 2001.

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

[3] 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.

[4] T. Kanda, H. Ishiguro, M. Imai, T. Ono, and K. Mase. A constrcutive approach for developing interactive humanoid robots. In *Proc. of IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pp. 1265–1270, 2002.

[5] M. Pelaez-Nogueras, J. L. Gewirtz, and M. M. Markham. Infant vocalizations are conditioned both by maternal imitation and motherese speech. *Infant behavior and development*, Vol. 19, p. 670, 1996.

[6] T. Kohonen. *Self-Organization and Assosiative Memory*. Springer-Verlag, New York, 1984.

[7] S. Amari. Neural theory of association and concept-formation. *Biological Cybernetics*, Vol. 26, pp. 175–185, 1977.

[8] P. Rubin and E. Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter Measuring and modeling speech production. Springer-Verlag, New York, 1998.

[9] N. Masataka. Effects of contingent and noncontingent maternal stimulation of vocal behavior of 3 to 4-month-old japanese infants. *J. of Child Language*, Vol. 20, pp. 303–312, 1993.