The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Phoneme Acquisition Model based on Vowel Imitation using Recurrent Neural Network

Hisashi Kanda, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani, and Hiroshi G. Okuno

Abstract - **A phoneme-acquisition system was developed using a computational model that explains the developmental process of human infants in the early period of acquiring language. There are two important findings in constructing an infant's acquisition of phonemes: (1) an infant's vowel like cooing tends to invoke utterances that are imitated by its caregiver, and (2) maternal imitation effectively reinforces infant vocalization. Therefore, we hypothesized that infants can acquire phonemes to imitate their caregivers' voices by trial and error, i. e., infants use self-vocalization experience to search for imitable and unimitable elements in their caregivers' voices. On the basis of this hypothesis, we constructed a phoneme acquisition process using interaction involving vowel imitation between a human and an infant model. Our infant model had a vocal tract system, called the Maeda model, and an auditory system implemented by using Mel-Frequency Cepstral Coefficients (MFCCs) through STRAIGHT analysis. We applied Recurrent Neural Network with Parametric Bias (RNNPB) to learn the experience of self-vocalization, to recognize the human voice, and to produce the sound imitated by the infant model. To evaluate imitable and unimitable sounds, we used the prediction error of the RNNPB model. The experimental results revealed that as imitation interactions were repeated, the formants of sounds imitated by our system moved closer to those of human voices, and our system could self-organize the same vowels in different continuous sounds. This suggests that our system can reflect the process of phoneme acquisition.**

I. Introduction

Our goal was to clarify how to acquire the ability to distinguish phonemes in the early development of human infants. Infants can acquire spoken language through imitating the vocal output of their parents. This ability is closely related to the cognitive development of language.

Developmental psychologists have demonstrated that an infant's vowel-like cooing tends to invoke utterances that are imitated by its caregiver's [1] and that maternal imitation effectively reinforces infant vocalization [2]. Infants have no innate knowledge of phonemes and regard a sound of phoneme sequences as continuous acoustic signals. As they grow, infants acquire the ability to discover phoneme units in continuous speech sounds by prosody, rhythm, stress, and whether they can imitate the sound or not.

We hypothesized that infants can acquire phonemes to imitate their caregiver's voices repeatedly by trial and error, i.e., **infants use self-vocalization experience to search for imitable and unimitable elements in their caregiver's voices.** We define phoneme acquisition in this paper as follows: Infants can produce sounds close to caregivers' voices.

The human-development studies have designed vocal imitation systems that duplicate the acquisition process of vowels [3]-[5]. The studies assume that acoustic signals consist of discrete phoneme sequences in advance, and they search for vocal-tract shapes corresponding to phonemes. However, articulatory movements for the same phoneme dynamically change according to the context of continuous speech (e.g. co-articulation). This effect derives from a physical constraint where articulation should be continuous in generating sounds. Recent neuroscience studies seem to support the relationship between articulations and voices as an active process involving motor cognition [6, 7].

This paper targets phoneme acquisition obtained by continuous-sound segmentation and imitation. We have already developed and verified a vocal-imitation system using Recurrent Neural Network with Parametric Bias (RNNPB) [8]. Neural network with recursive connections can model continuous utterances of infant as "dynamics." In this paper, we represented a infant model as a vocal-tract system, called the Maeda model [9]. To acquire phonemes, the infant model repeated learning and producing imitable sounds which the model selected from human voices. Our infant model used the prediction error of RNNPB to evaluate imitable and unimitable sounds produced by humans. STRAIGHT analysis is a kind of pitch analysis depending on the fundamental frequency (F0) of the sound [10]. As a result of the eliminating F0 of acoustic parameters, the analysis decreases the difference in sounds produced by humans and the Maeda model. The segmenting method using RNNPB can divide several kinds of sequences into primitive sections which are encoded as a set of parameters, called PB values [11]. We expected that our infant model could manipulate the encoded phonemes to imitate human voices.

Section II gives an overview of our phoneme-acquisition process, and it describes the infant and the RNN models. Section III describes our imitation model and the system. Section IV presents the experimental results for vowel acquisition. Section V discusses the vowels acquired and imitated with our system, and Section VI concludes the paper.

II. Phoneme Acquisition System

A. Overview

As we can see from Fig. 1, our phoneme-acquisition process consists of four phases: learning, recognition, generation, and selection.

H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno is with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan {hkanda, ogata, tall, komatani, okuno}@kuis.kyoto-u.ac.jp

Phase 1: Learning (Experiencing self-vocalization)

The phoneme-acquisition system produces sounds, and makes a connection between an articulatory movement and the sound produced by the movement.

Phase 2: Recognition (Hearing parent's sounds)

We enter voices into the system. The system recognizes the voices with an articulation producing the same dynamics as in the heard voice.

Phase 3: Generation (Imitating sounds)

The system uses the articulation to imitate the voice.

Phase 4: Selection (Exploring imitable sounds)

The system calculates the error between the heard and imitated sounds. The errors in imitable sounds are small and those in unimitable ones are large.

The process corresponds to the babbling and imitation of vowels in 3-6-month-old infants [12]. Our system repeats the process to acquire phonemes, especially vowels. Our model can self-organize to connect an articulatory movement with the corresponding sound dynamics. Additionally, the connection is available in the recognition and generation phases to imitate human voices.

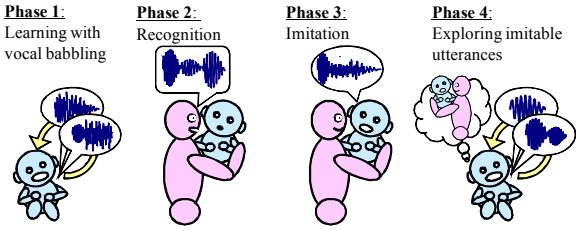


Figure 1 Phoneme acquisition process.

B. Infant Model

1) Auditory System: We used a kind of Mel-Frequency Cepstrum Coefficients (MFCCs) called STRAIGHT MFCCs, which were obtained from the power spectrum of a sound waveform segment. In STRAGHT MFCC, the power spectrum was calculated by using STRAIGHT instead of short term Fourier transform of its segment. STRAIGHT analysis is a kind of pitch analysis in which the window length is set depending on the F0 of the sound. The power spectrum experiences no interference caused by the F0 of the vocal source. The MFCCs are calculated by taking the discrete cosine transform of mel-scaled log filter bank energies.

2) Vocal-tract System: We used the physical vocal-tract model proposed by Maeda [9]. This model has seven parameters determining the vocal tract shape, which were derived by principal components analysis of cineradiographic and labiofilm data from French speakers. Table I lists the seven parameters, and Table II has the first and second formant (F1 and F2) of vowels produced by the Maeda model.

A speech production model simulating the human vocal tract system incorporates the physical constraints of the articulatory mechanism. The parameters of the vocal tract with physical constraints are better for continuous-speech synthesis than acoustic parameters such as the sound spectrum. This is because the temporal change in the vocal-tract parameters is continuous and smooth, while that of the acoustic parameters is complex, and it is difficult to interpolate the latter parameters between phonemes. Although there are other vocoders, such as PARCOR [13] and STRAIGHT, we think that the Maeda model is the most appropriate to simulate the developmental process of infant's speech. This is because it has physical constraints based on anatomical findings. This model for generating acoustic signals is a very simplified articulatory model, and the sound units corresponding to phonemes are expressed in these articulatory terms.

Table I Parameters of Maeda model

Number	Name	Value: +3 … -3
1	Jaw position (JP)	Upper <-> Lower
2	Tongue dorsal position (TDP)	Forward <-> Back
3	Tongue dorsal shape (TDS)	Upper <-> Lower
4	Tongue tip position (TTP)	Upper <-> Lower
5	Lip opening (LO)	Closed <-> Open
6	Lip protrusion (LPR)	Forward <-> Back
7	Larynx position (LP)	Upper <-> Lower

Table II Average vowel formants of Maeda model

	/a/	/i/	/u/	/e/	/o/
F1 (Hz)	667	234	269	401	500
F2 (Hz9	1214	2161	924	1894	902

C. Learning Algorithm

This subsection describes the method we used to learn and segment temporal-sequence dynamics. We applied the RNNPB model, which was first proposed by Tani and Ito [14] as a forwarding forward model. It generates complex movement sequences, which are encoded as the limit-cycling dynamics and/or fixed-point dynamics of RNN.

1) RNNPB model: The RNNPB model has the same architecture as the conventional Jordan-type RNN model [15], except for the PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a constant value throughout each temporal sequence and are used to implement mapping between fixed-length values and temporal sequences. Figure 2 outlines the network configuration for the RNNPB model. Unlike the Jordan-type RNN model, the RNNPB self organizes the values in the PB nodes that encode the sequence during the learning process. The common structural properties of the training-data sequences are acquired as connection weights by using the back-propagation through time (BPTT) algorithm [16], as in a conventional RNN. The specific properties of each individual temporal sequence are simultaneously encoded as PB values. As a result, the RNNPB model self-organizes mapping between the PB values and the temporal sequences.

2) Segmenting Temporal-sequence Data: Our segmenting method determines the segmentation boundaries using the prediction error in the RNNPB model. Systems using this approach usually consist of dynamic recognizers that predict the target sequences. The dynamic sequence is articulated based on how predictable the recognizer is. The method we used to segment acoustic signals with articulatory movements uses the prediction error in the RNNPB model and the number of segmentations. Its description is as follows. Consider the problem of segmenting a dynamic sequence,

$D(t)$, whose length is T into N sections, which are represented as S_i (i = 0, ⋯ , N−1). The boundary step between S_i−1 and S_i is represented by $t = s_i$, i.e., S_i is defined as $[s_i, s_i+1]$. The segmenting process consists of five steps.

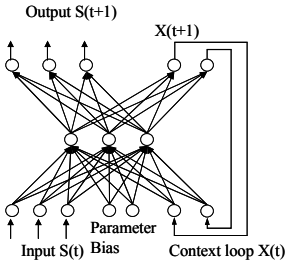


Figure 2 Network configuration of RNNPB model

Step 1) Initialization: The given sequence is divided into N sections. Each section has the same length. The boundary step, s_i (i= 0, ⋯, N), is set as $T \cdot i / N$.
Step 2) RNNPB training: The connection weights and PB values of the RNNPB model are updated with the given sequence, while the PB values are kept constant in each section, S_i.
Step 3) Calculate prediction errors: The prediction sequences of the RNNPB model, $P(t)$, are calculated in each S_i by using forward calculation, and the average prediction errors, E_i, are obtained.
Step 4) Update length of each section: The boundary step, s_i (i = 1, ⋯, N −1), is updated by using the following rules:

$$s_{i+1} \leftarrow \begin{cases} s_{i+1} - ds & if\ E_i \geq E_{i+1} \\ s_{i+1} + ds & if\ E_i < E_{i+1} \end{cases} \quad (1)$$

where d_s is a parameter to update the section length.
Step 5) Repeat Steps 2 to 4 until entire error is converged.

If a sequence is generated by using simple dynamics, the prediction error in the RNNPB will be small, even when the PB values are fixed. However, if a sequence is generated by using multiple dynamics, the prediction error at the boundary between dynamics will increase. The algorithm can decrease the error by modifying the position of each boundary.

3) Learning PB Vectors: The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The step length of the ith section S_i in a sequence is denoted by s_i+1 − s_i. For each of the articulatory and sound parameters outputs, back propagated errors with respect to PB nodes are accumulated and used to update PB values. The update equations for the kth unit of the PB nodes at section S_i are

$$\delta\rho_i = \varepsilon \sum_{t=0}^{T} \delta_t^{bp^i} \quad (2)$$

$$p_t = sigmoid(\rho_t / \zeta) \quad (3)$$

In Eq. (2), δ_t^{bp} represents the delta error back propagated from the output nodes to the PB nodes and is integrated over period T steps. Internal value ρ_t is updated using the delta force, as shown in Eq. (3). The ε and ζ are learning coefficients. It is integrated over the period from s_i to s_i+1 steps. Then, the current PB values, p_{ik}, are obtained from the sigmoidal outputs of the updated internal values in Eq. 1.

D. Calculation in Recognition and Generation Phases

After the RNNPB model is organized in the learning phase, it is used in the recognition and generation phases. The recognition phase corresponds to how infants recognize sounds presented by parents, i.e., to how the PB values are obtained. The PB values of each section are calculated from Eqs.2 and 3 by using the organized RNNPB without updating the connection weights. The boundary steps of each sequence are determined by the prediction errors in the organized RNNPB. However, there are no articulatory data because the system is only hearing sounds without articulating them, unlike in the learning phase. The initial vocal tract values (these are all zeros) are input to the vocal tract units of the input layer in Step 0, and the outputs are calculated forward in the closed-loop mode from Step 1. More generally, the outputs in the articulatory output layer in Step t−1 are the input data in the articulatory input layer in Step t. This calculation is called a *closed loop calculation*. The generation phase corresponds to what articulation values have been calculated. The articulatory output of the RNNPB model is obtained in a *closed loop calculation*. The PB values obtained in the recognition phase are input to the RNNPB in each step

III. Phoneme Acquisition System

A. Experimental System

Our experimental system is illustrated Fig. 3. We targeted vowel-sound segmentation and imitation in this paper. Our system does not know the numbers and kinds of vowels in sounds. This condition corresponds to human infants who do not have knowledge or skills to deal with phonemes. The infant model learns self-vocalization in the learning phase. In the first learning, we used a cubic interpolation method to produce articulatory parameters (APs) for the Maeda model. In the second or later learning, the model uses APs corresponding to imitated sounds in the selection phase. Then, the Maeda model uses the APs to produce sounds, which are then transformed into MFCCs by the auditory system described in subsection II-B.1. Finally, the RNNPB model learns each of the MFCC and AP sequences, which are normalized and synchronized. Parameter ds is set at 0.1. In the recognition phase, the infant model listens to human voices. MFCC sequences of vowels sound produced by a human are entered into the organized RNNPB model. The RNNPB model calculates the corresponding PB values for the given sequence to associate the articulatory movements with the sounds.

In the generation phase, the infant model generates imitated sounds. The organized RNNPB produces articulatory sequences using the PB values obtained in the recognition phase. Then, the sequences are input into the Maeda model to

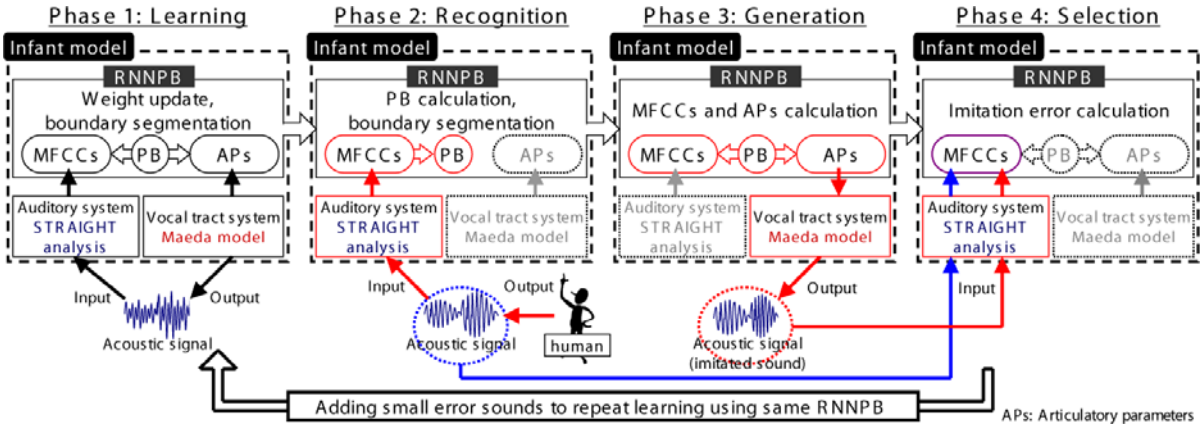


Fig. 3. Diagram of experimental system

produce sounds.

In the selection phase, the infant model discriminates re-learning sounds. The organized RNNPB calculates all MFCC errors in the sound between humans and the infant model. Then, the model selects the re-learning sound whose MFCC error is less than average error of the imitated sounds in the generation phase.

B. Sound Parameter: MFCCs

The acoustic signals in our experiment were single channel with a sampling frequency of 10 kHz. We set the number of filter banks to 12. We formed five-dimensional vectors from the low-third to the low-seventh dimension out of 12-dimensional MFCC vectors by using STRAIGHT analysis. The vectors produced from speech sounds remained vowel features, and they were almost independent of speakers.

C. Articulatory Parameters: Maeda Model Parameters

We used the Maeda parameters in Table I except for the seventh parameter LP. This is because, when the Maeda model produces vowel sounds, the LP is steady. In the generation phase, it was possible for the Maeda parameters produced by the RNNPB to temporally fluctuate without human physical constraints. This occurred if the system did not easily associate the articulatory movements of an inexperienced sound. Therefore, to prevent extraordinary articulation, we temporally smoothed the Maeda parameters produced by the RNNPB. Concretely, the parameters in each step were calculated by averaging those of the adjacent steps.

IV. Vowel Acquisition Experiment

We carried out two experiments. In the first, we examined the effect of self-vocalization in the initial learning phase. In the second, we examined the phoneme-acquisition capabilities of our infant model through our process.

A. Experiment 1: Random Babbling

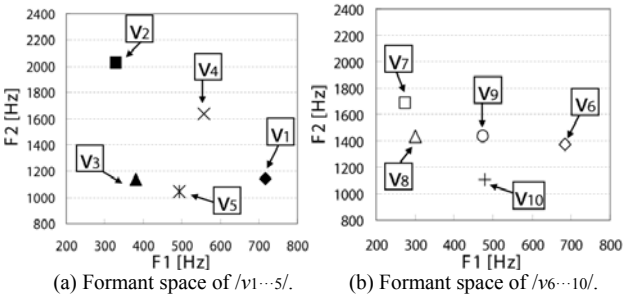
Our infant model learned random babbling in the initial learning phase. Random babbling meant that our model used vowel-like sounds produced by random articulation of the Maeda model to learn self-vocalization. We used the 10 kinds of random sounds /v_1/, ⋯, /v_{10}/ in Fig. 4 to create two sets of learning patterns consisting of three sounds (each pattern was 45 steps at 30ms/step) as follows.

Set-1: /$v_1v_2v_3$/, /$v_2v_3v_4$/, /$v_3v_4v_5$/, /$v_4v_5v_1$/, /$v_5v_1v_2$/, /$v_2v_1v_5$/, /$v_1v_5v_4$/, /$v_5v_4v_3$/, /$v_4v_3v_2$/, /$v_3v_2v_1$/,

Set-2: /$v_6v_7v_8$/, /$v_7v_8v_9$/, /$v_8v_9v_{10}$/, /$v_9v_{10}v_6$/, /$v_{10}v_6v_7$/, /$v_7v_6v_{10}$/, /$v_6v_{10}v_9$/, /$v_{10}v_9v_8$/, /$v_9v_8v_7$/, /$v_8v_7v_6$/.

Set-1 and Set-2 has the different distribution of the formant: Set-1 corresponded to large articulatory movements, and Set-2 corresponded to limited movements.

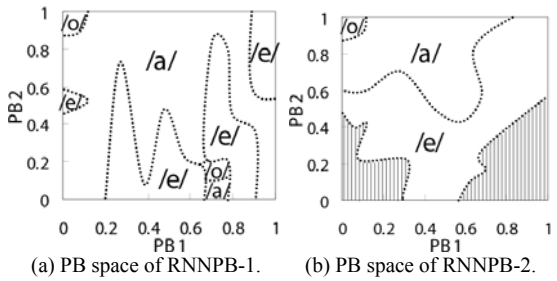
We used two RNNPBs: the first learning Set-1 was called RNNPB-1, and the second learning Set-2 was called RNNPB-2. The organizations of Both RNNPBs were organized as follows: 11 input/output nodes, 40 hidden nodes, 5 context nodes, and 2 PB nodes. Each RNNPB had 200,000 learning iterations where $ds = 0.1$ and $N = 8$.



(a) Formant space of /$v_{1\cdots5}$/. (b) Formant space of /$v_{6\cdots10}$/.

Fig. 4. Formant space of initial learning set.

Figure 5 shows the analysis of PB space for both organized RNNPBs. This analysis was conducted in five steps: 1) The PB space was divided into 10 x 10 lattices. 2) We used APs obtained through a *closed loop calculation* to produce a 300ms sound for all lattices. 3) The F1 and F2 averages of the second half of all produced sounds were calculated. 4) The square error of F1 and F2 averages from those in Table II were calculated for all vowels. 5) A vowel corresponding to the minimum square error was placed at each lattice point. In Figure 5, all vowels have a nonlinear distribution for the F1 and F2 formants. The vowels /*a*/, /*e*/ are very widely distributed in the PB space. The PB values in the shaded areas of Fig. 5(b) could not produce sounds. We found that RNNPB-1 could produce continuous sounds, but that RNNPB-2 could not.



(a) PB space of RNNPB-1. (b) PB space of RNNPB-2.
Fig. 5. Results of PB analysis in initial learning phase.

B. Experiment 2: The Simulation of Phoneme Acquisition

We carried out the simulations of vowel acquisition using two RNNPBs. The infant model repeated our phoneme acquisition process three times. In the first learning phase, the learning conditions were the same as for RNNPB-1 in the 1st experiment. In the recognition phase, RNNPB-3 recognized the three-vowel sounds of Speaker-1, and RNNPB-4 recognized those of Speaker-2 listed in Table IV (each sound was 1350ms). Table III shows the average formants for each speaker. We set the segmentation number, *N*, as the least MFCC error in all sounds obtained by each organized RNNPB. In the generation phase, we used the PB values and the boundary steps to reproduce each of the recorded sounds. In the selection phase, the infant model selected sounds where MFCC errors were less than the average error in recognizing and generating sounds. In the second or later learning phases, each organized RNNPB relearned the selected sounds in the selection phase. There were 100,000 iterations for learning. Figure 6 has a bar chart of the average imitation errors in the 1st, 2nd, and 3rd generation phases. The error in each RNNPB was reduced by repeating the proposed process. Figure 7 shows the PB space analysis of RNNPB-3 and 4 in the 2nd and 3rd learning phases. Compared with Fig. 5(a), there are clear vowel distributions in the 2nd and 3rd PB spaces for RNNPB-3 and 4. Furthermore, a new vowel appeared in the 3rd PB space: /*i*/ for RNNPB-3, and /*u*/ for RNNPB-4.

Table III Average vowel formants (top: Speaker-1, bottom: Speaker-2).

	/a/	/i/	/u/	/e/	/o/
F1 (Hz)	667	234	269	401	500
F2 (Hz9	1214	2161	924	1894	902
	/a/	/i/	/u/	/e/	/o/
F1 (Hz)	667	234	269	401	500
F2 (Hz9	1214	2161	924	1894	902

Table IV Input sounds in recognition phase.

/aeo/ /aeu/ /aia/ /aie/ /aio/ /aiu/ /aoa/ /aou/ /aue/
/eai/ /eia/ /eiu/ /eoa/ /eoe/ /eoi/ /eou/ /eua/ /eue/
/iae/ /iai/ /ieo/ /ioa/ /ioe/ /iua/ /iue/ /iui/ /iuo/
/oae/ /oai/ /oao/ /oau/ /oei/ /oeo/ /oiu/ /oue/ /oui/
/uai/ /uao/ /uea/ /uei/ /ueo/ /ueu/ /uio/ /uiu/ /uoa/

Figure 9 shows the formant space of imitated sounds for RNNPB-3 in the 1st, 2nd, and 3rd generation phase. The phonemes in the set of three-vowel data in Table IV were aligned to the length of the three longest sections for each imitated sound. We fitted normal distributions to Speaker-1's vowel formants as colored ellipses, and those of imitated sounds' formants representing vowels as gray-scale ellipses. After repeating our process, each vowel in the imitated sounds except for /*a*/ gradually became closer to Speaker-1's vowel formants. We achieved the same results for RNNPB-4. We confirmed that our model could imitate vocal sounds involving arbitrary numbers of vowels using the vowel space in the RNNPB. The space was acquired by "babbling" with the vocal-tract model with only a few sets of vowel sounds.

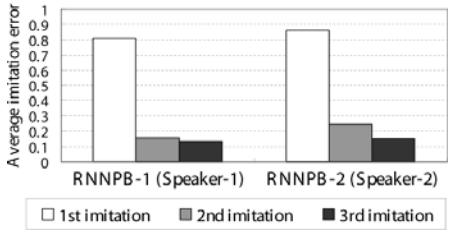
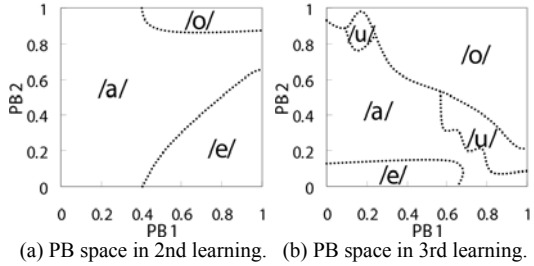
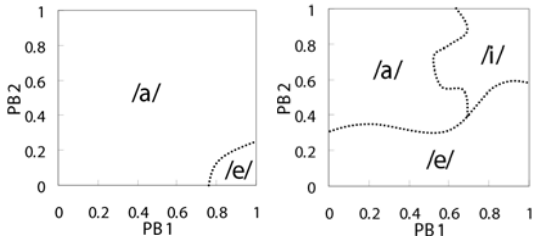


Fig. 6. Average imitation error in 1st, 2nd, and 3rd generation phases.



(a) PB space in 2nd learning. (b) PB space in 3rd learning.
Fig. 7. PB space for RNNPB-3



(a) PB space in 2nd learning. (b) PB space in 3rd learning.
Fig. 8. PB space for RNNPB-4

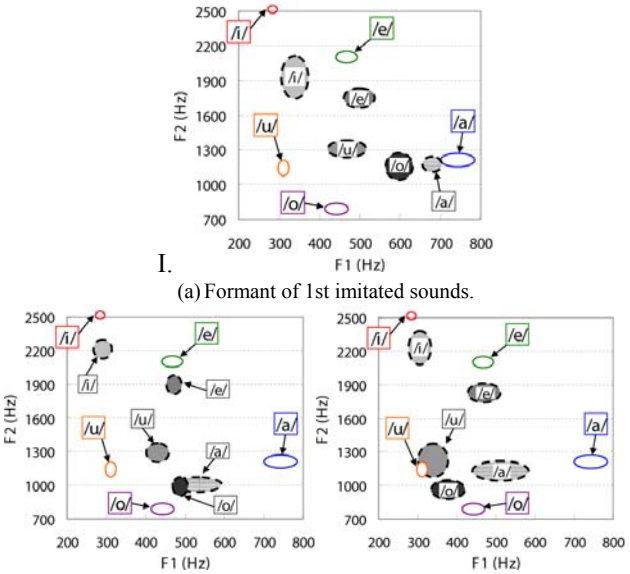
V. Discussion

A. Articulation to Imitate Sounds

Experiment 1 revealed the change in imitation capabilities under conditions with different articulatory movements. Figure 4 and 5 confirmed that large articulatory movements helped our infant model to produce many kinds of sounds. There were especially large differences between maximum and minimum values of JP, TDP, and LO in Table I.

In fact, these result corresponded to infant babbling. Jakobson demonstrated that infants could produce sounds by maximum and/or minimum movements in articulation in the early period of babbling [17]. Furthermore, the development of controllability for forward and back tongue movements delayed that for up and down tongue movements [18]. The TDP movement of the Maeda model had a close relation to forward and back tongue movements. The results of the experiment suggested that one of the necessary conditions to be able to imitate sounds was extreme articulation for forward and back tongue movements that delayed those for up

and down tongue movements in infants.



I.

(a) Formant of 1st imitated sounds.

(b) Formant of 2nd imitated sounds. (c) Formant of 3rd imitated sounds.

Fig. 9. Formant space in generation phase for RNNPB-3

B. Vowel Acquisition

Our system could improve imitated sounds close to the formants of speakers' voices. The reason of failure to imitate vowel /*o*/ was presumably because human infants have difficulties producing this vowel. Actually, there were large overlaps between the distribution of vowel /*o*/ and the others in F1-F2 space for two Japanese infants [19]. This suggests that our model reflects the process of vowel acquisition.

However, the formants for vowel /*a*/ in the imitated sounds gradually became dissimilar to those of speakers' voices. The reason for this is that vowel /*a*/ was overtraining for all RNNPBs in the 2nd experiment. It was not impossible for RNNPB-3 and 4 to produce vowel /*a*/. In Figs. 7 and 8, there was a large area for vowel /*a*/ in each PB space. The error in imitating sounds including vowel /*a*/ tended to be less than when other vowels were included. This is because RNNPB re-learned vowel /*a*/ numerous times. We need to take into consideration error in the selection phase.

VI. Conclusions and Future Work

We developed a phoneme-acquisition system based on the interaction in caregiver-infant vocal imitation consisting of four phases: learning, recognition, generation, and selection. Our infant model inputs sounds through STRAIGHT analysis, and outputs sounds through the Maeda model. Using self-vocalization experience, the model evaluates imitable and unimitable sounds produced by humans. Through experiments, we confirmed that many articulatory movements helped the infant model to imitate speakers' sounds, and that our process enabled the model to acquire the phonemes of speakers by the definition in Section I. Furthermore, the order of vowels acquired by our system corresponded to that by real infants. As a result, we confirmed the accuracy of simulations of our phoneme-acquisition process. Our future work includes the use of smaller vocal-tract model of infant and the imitation of "consonants" through simulating caregiver-infant interactions.

Acknowledgement

This research was partially supported by Grant-in-Aid for Scientific Research (S) 19100003, Creative Scientific Research 19GS0208. Global COE, and RIKEN.

References

[1] 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.

[2] M. Pel´aez-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.

[3] B. de Boer, "Self-organization in vowel systems," *J. Phonetics*, vol. 28, no. 4, pp. 441–465, 2000.

[4] P. Y. Oudeyer, "The self-organization of speech sounds," *J. Theoretical Biology*, vol. 233, no. 3, pp. 435–449, 2005.

[5] K. Miura and et al., "Vowel acquisition based on visual and auditorymutual imitation in mother-infant interaction," in *ICDL2006*, 2006.

[6] L. Fadiga and et al., "Speech listening specifically modulates the excitability of tongue muscles: a tms study," *Euro. J. Cogn. Neurosci.*, vol. 15, pp. 399–402, 2002.

[7] G. Hickok, B. Buchsbaum, C. Humphries, and T. Muftuler, "Auditorymotor interaction revealed by fmri," *Area Spt. J. Cogn. Neurosci.*, vol. 15, no. 5, pp. 673–682, 2003.

[8] H. Kanda, T. Ogata, K. Komatani, and H. G. Okuno, "Segmenting acousitic signal with articulatory movement using recurrent neural network for phoneme acquisition," in *IEEE/RSJ IROS-2008*, 2008.

[9] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," *Speech production and speech modeling*, pp. 131–149, 1990.

[10] H. Kawahara, K. Masuda, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[11] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Twoway translation of compound sentences and arm motions by recurrent neural networks," in *IEEE/RSJ IROS-2007*, 2007.

[12] P. K. Kuhl and et al., "Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)," *Philos. Trans. R. Soc. B: Biol. Sci.*, vol. 363, no. 1493, pp. 979–1000, 2008.

[13] N. Kitawaki, F. Itakura, and S. Saito, "Optimum coding of transmission parameters in parcor speech analysis synthesis system," *Trans. IEICE Japan*, vol. J61-A, no. 2, pp. 119–126, 1978.

[14] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Trans. On SMC Part A*, vol. 33, no. 4, pp. 481–488, 2003.

[15] M. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," in *Annu. Conf. Cog. Sci. Soc.*, 1986, pp. 513–546.

[16] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation*. MIT Press, 1986.

[17] R. Jakobson, "Why 'mama' and 'papa'?" in *Perspectives in psychological theory*. International Universities Press, 1960.

[18] B. de Boysson-Bardies, "Ontogeny of language-specific syllabic production," in *Developmental neurocognition: Speech and face processing in the first year of life*. Kluwer Academic Publishers, 1993.

[19] K. Ishizuka and et al., "Longitudinal developmental changes in spectral peaks of vowels produced by japanese infants," *J. Acoust. Soc. Am.*,vol. 121, no. 4, pp. 2272–2282, 2007.