2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Continuous Vocal Imitation with Self-organized Vowel Spaces in Recurrent Neural Network

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

Abstract— **A continuous vocal imitation system was developed using a computational model that explains the process of phoneme acquisition by infants. Human infants perceive speech sounds not as discrete phoneme sequences but as continuous acoustic signals. One of critical problems in phoneme acquisition is the design for segmenting these continuous speech sounds. The key idea to solve this problem is that articulatory mechanisms such as the vocal tract help human beings to perceive speech sound units corresponding to phonemes. To segment acoustic signal with articulatory movement, we apply the segmenting method to our system by Recurrent Neural Network with Parametric Bias (RNNPB). This method determines the multiple segmentation boundaries in a temporal sequence using the prediction error of the RNNPB model, and the PB values obtained by the method can be encoded as kind of phonemes. Our system was implemented by using a physical vocal tract model, called the Maeda model. Experimental results demonstrated that our system can self-organize the same phonemes in different continuous sounds, and can imitate vocal sound involving arbitrary numbers of vowels using the vowel space in the RNNPB. This suggests that our model reflects the process of phoneme acquisition.**

I. INTRODUCTION

Our goal is to clarify how to acquire the ability to distinguish phonemes in the early period of human infants. Human infants can acquire spoken language through vocal imitation of their parents. Despite their immature bodies, they can imitate their parents' speech sounds by generating those sounds repeatedly by trial and error. This ability is closely related to the cognitive development of language.

Many researchers took notice of the relationship between articulatory movements and sounds produced by the movements. They have designed vocal imitation systems that duplicate the developmental process of infants' vowel acquisition [1], [2], [3]. These studies were based on the idea that articulatory mechanisms such as the vocal tract enable us to acquire phonemes, i.e. sounds in the form of phonemes are characterized by motor articulations. This idea has been advocated as the *motor theory of speech perception* [4], and recent neuroscience studies seem to show the idea to be an active process involving motor cognition [5], [6].

Segmenting acoustic signals with articulatory movements is essential for vocal imitation and phoneme acquisition; the reason is that human infants do not know the given phonetic distinction inherently. The human development studies described above 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. coarticulation). This effect derives from a physical constraint that articulatory movements should be continuous in sound generation. We assume that human infants regard phoneme sequences as continuous acoustic signals. As they grow, infants will acquire the ability to discover phoneme units in a continuous speech sound by prosody, rhythm, stress and whether they can imitate the sound or not.

We use Recurrent Neural Network with Parametric Bias (RNNPB) [7] to segment a temporal sequence consisting of acoustic signal with articulatory movement. From the view point of considering sounds as temporal sequences, we have already developed a vocal imitation system using a phisical vocal tract model, called the Maeda model [8], and verified the segmentation ability [9]. In this paper, we target vocal imitation by continuous sound segmentation. We apply the segmenting method by RNNPB [10] to our imitation system and use the acoustic parameters calculated by the STRAIGHT analysis [11]. The segmenting method can divide several kinds of sequences into primitive sections using the prediction error of RNNPB. The primitives are encode as a set of parameters, called PB values. The STRAIGHT analysis is a kind of pitch analysis depending on the fundamental frequency (F0) of the sound. As a result of eliminating F0 of the acoustic parameters, the analysis decreases the difference of produced sounds between humans and the Maeda model. Due to above-mentioned two approaches, it is expected that our imitation system can manipulate the encoded phonemes to imitate heard sounds.

Section II gives an overview of our imitation process, and it describes the vocal tract model and the RNN model. Section III describes our imitation model and the system. Section IV gives the experimental results of continuous vocal imitation. Section V discusses the vowel acquisition and imitation of our system, and Section VI concludes the paper.

II. VOCAL IMITATION PROCESS AND MODEL

A. Overview of Our Imitation Process

In this section, we present an overview of our imitation system. As illustrated in Fig. 1, our imitation process consists

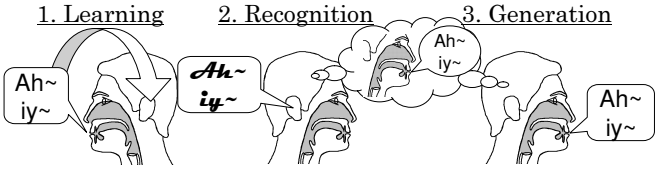

Fig. 1. Imitation process.

H. Kanda, T. Ogata, T. Takahashi, K. Komatani and H. G. Okuno are 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

of three phases: learning, recognition, and generation.

1) Learning (Babbling)
   The vowel imitation system makes articulatory movements to produce sounds, and it makes a connection between an articulatory movement and the sound produced by the movement. This phase corresponds to babbling in infants.
2) Recognition (Hearing parents' speech sounds)
   In this phase, we put a speech sound into the system. The system recognizes the sounds with an articulation producing the same dynamics as the heard sound.
3) Generation (Vocally imitating heard sounds)
   Finally, the system uses the articulatory movement to imitate a speech sound.

The imitation process corresponds to the babbling and vowel imitaion in 3-6-month-old infants [12]. The learning phase uses the RNNPB method of segmenting temporal sequences. Our model can self-organize so as to connect an articulatory movement with the corresponding sound dynamics. Additionally, in the recognition and generation phases, the connection is available to imitate speech sounds.

B. Physical Vocal Tract Model

A speech production model simulating the human vocal tract system incorporates the physical constraints of the articulatory mechanism and the acoustic constraints of speech production. 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 of 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.

We used the vocal tract model proposed by Maeda [8]. This model has seven parameters determining the vocal tract shape, and they were derived by principal components analysis of cineradiographic and labiofilm data from French speakers. Table I lists the seven vocal tract parameters. Although there are other vocoders, such as PARCOR [13] and STRAIGHT [11], we think that the Maeda model is the most appropriate to simulate the developmental process of infant's speech. Because the Maeda model 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.

Each Maeda parameter takes on a real value between -3 and 3, and may be regarded as a coefficient weighting an eigenvector. The sum of these weighted eigenvectors is a vector of points in the midsagittal plane that defines the outline of the vocal tract shape. The resulting vocal tract shape is transformed into an area function, which is processed to obtain the acoustic output and spectral properties of the vocal tract during speech.

TABLE I
PARAMETERS OF THE MAEDA MODEL.

Parameter number	Parameter name
1	Jaw position (JP)
2	Tongue dorsal position (TDP)
3	Tongue dorsal shape (TDS)
4	Tongue tip position (TTP)
5	Lip opening (LO)
6	Lip protrusion (LPR)
7	Larynx position (LP)

C. Learning Algorithm

This subsection describes the method to learn and segment temporal sequence dynamics. We apply the RNNPB model, which was first proposed by Tani [7] as the forwarding forward model. It generates complex movement sequences, which are encoded as the limit-cycling dynamics and/or the fixed-point dynamics of the RNN.

1) RNNPB model: The RNNPB model has the same architecture as the conventional Jordan-type RNN model [14], 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 a mapping between fixed-length values and temporal sequences. Figure 2 shows the network configuration of the RNNPB model.

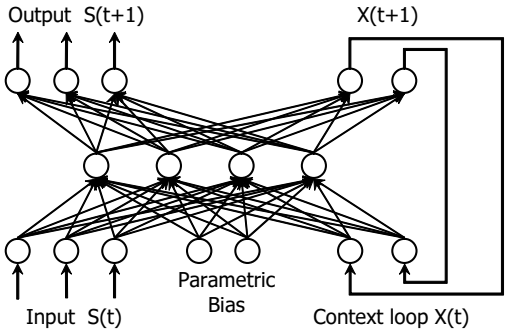

Fig. 2. RNNPB model.

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 backpropagation through time (BPTT) algorithm [15], as in a conventional RNN. Meanwhile, the specific properties of each individual temporal sequence are simultaneously encoded as PB values. As a result, the RNNPB model self-organizes a 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 of the RNNPB model. Systems using this approach usually consist of dynamic recognizers that predict the target sequences. The dynamic sequence is articulated based on the predictability of the recognizer. The method we used to segment acoustic signals with articulatory movements uses the prediction error of 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 $(i = 0, \cdots, N-1)$. The boundary step between S_{i-1} and S_i is represented by $t = s_i$, that is, S_i is defined as $[s_i, s_{i+1}]$. The segmenting process consists of five steps.

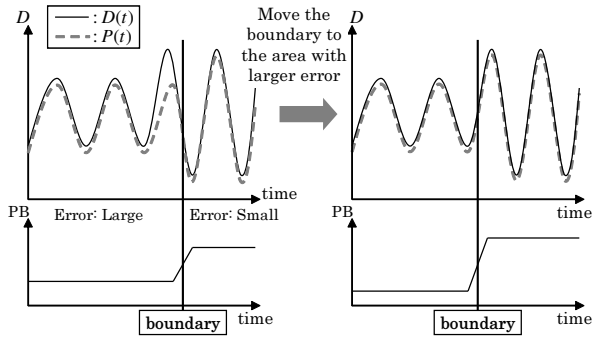

Fig. 3. Segmenting multiple dynamics.

Step 1: Initialization

The given sequence is divided into N sections. Each section has the same length. The boundary step s_i ($i = 0, \cdots, N$) is set as follows.

$$s_i \longleftarrow i \cdot T/N \tag{1}$$

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

In each S_i, the prediction sequences of the RNNPB model, $P(t)$, are calculated, and the average prediction errors, E_i, is obtained as follows.

$$E_i \longleftarrow \frac{1}{s_{i+1} - s_i} \cdot \sum_{t \in S_i} ||D(t) - P(t)|| \tag{2}$$

Step 4: Update the length of each section

The boundary step s_i ($i = 1, \cdots, N-1$) is updated by using the following rules:

$$s_i \longleftarrow \begin{cases} s_i - ds & if\ E_{i-1} \geq E_i \\ s_i + ds & if\ E_{i-1} \leq E_i, \end{cases} \tag{3}$$

where ds is a parameter to update the section length.

Step 5: Repeat Steps 2 to 4 until the whole error is less than the threshold.

If a sequence is generated by using simple dynamics, the prediction error of 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 as shown in Fig. 3. The algorithm can decrease the error by modifying the position of each boundary.

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

$$\delta\rho_{i,k} = \varepsilon \cdot \sum_{t=s_i}^{s_{i+1}} \delta_{i,k}(t), \tag{4}$$

$$p_{i,k} = sigmoid(\rho_{i,k} + \delta\rho_{i,k}), \tag{5}$$

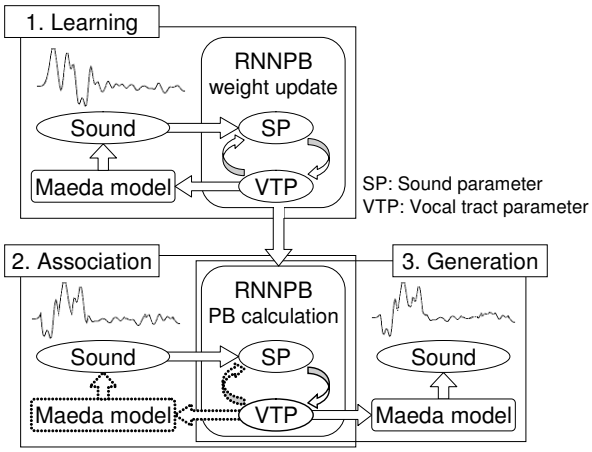

Fig. 4. Diagram of the experimental system.

where ε is a coefficient. In Eq. 4, the δ force for updating the internal values of the PB $\rho_{i,k}$ is obtained from the sum of the delta errors $\delta_{i,k}$. The delta error $\delta_{i,k}$ is backpropagated from the output nodes to the PB nodes: it is integrated over the period from s_i to s_{i+1} steps. Then, the current PB values $p_{i,k}$ are obtained from the sigmoidal outputs of the updated internal values in Eq. 5.

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 Eq. 4 and 5 by using the organized RNNPB without updating the connection weights. The boundary steps of each sequence are determined by the prediction errors of the organized RNNPB. However, there is no articulatory data because the system is only hearing sounds without articulating them, unlike in the learning phase. The initial vocal tract values (there are all zero) are input to 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 *closed loop calculation*.

The generation phase corresponds to what articulation values are 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. VOCAL IMITATION SYSTEM

A. Experimental System

Our experimental system is illustrated Fig. 4. In this paper, we target vowel sound segmentation and imitation. In order to simplify and prove the imitation ability of our system, we used a simple vocal tract model to make learning data consisting of the sequence of explicit vowels. Our system, however, does not know the number and kinds of vowels in sounds. This condition corresponds that human infants do not have knowledge and skills to deal with phonemes.

TABLE II
INPUT SOUND DATA IN THE RECOGNITION PHASE.

two-vowel		three-vowel					four-vowel
/ae/	/io/	/aeo/	/eai/	/iae/	/oae/	/uai/	/aiue/
/ai/	/iu/	/aeu/	/eia/	/iai/	/oai/	/uao/	/eoai/
/ao/	/oa/	/aia/	/eiu/	/ieo/	/oao/	/uea/	/iueo/
/au/	/oe/	/aie/	/eoa/	/ioa/	/oau/	/uei/	/oaiu/
/ea/	/oi/	/aio/	/eoe/	/ioe/	/oei/	/ueo/	/ueoa/
/ei/	/ou/	/aiu/	/eoi/	/iua/	/oeo/	/ueu/	
/eo/	/ua/	/aoa/	/eou/	/iue/	/oiu/	/uio/	
/eu/	/ue/	/aou/	/eua/	/iui/	/oue/	/uiu/	
/ia/	/ui/	/aue/	/eue/	/iuo/	/oui/	/uoa/	
/ie/	/uo/						

In the learning phase, we first use a cubic interpolation method to produce sequences of vocal tract parameters for the Maeda model as articulatory movements. Second, the sequences are put into the Maeda model to produce the corresponding sounds, which are then transformed into temporal sound parameters. Finally, the RNNPB learns each the sound and the vocal tract parameters, which are normalized and synchronized. In this phase, the parameter ds was set at 0.1. The size of the RNNPB model and the time interval of the sequence data differed according to the experiment.

In the recognition phase, speech sound data is put into the system. The corresponding PB values are calculated for the given sequence by the organized RNNPB in order to associate the articulatory movement with the sound data.

In the generation phase, the system generates imitation sounds by inputting the PB values obtained in the recognition phase into the organized RNNPB.

B. Sound Parameter

In this paper, we use a kind of Mel-Frequency Cepstrum Coefficients (MFCCs) as sound parameter, which are obtained from power spectrum of sound waveform segment. The power spectrum is calculated by STRAIGHT analysis instead of short-term Fourier transform of the segment. STRAIGHT analysis is a kind of pitch analysis in which the window length in the analysis is set depending on the F0 of the sound. The power spectrum has no interference caused by F0 of vocal source. The MFCCs are calculated by taking the discrete cosine transform of mel-scaled log filterbank energies.

In our experiment, speech signals were single channel with a sampling frequency 10 kHz. The number of filterbank was set to 12. We formed 5-dimensional vectors from low-third to low-seventh dimension out of 12-dimensional MFCC vectors. The vectors produced from speech sounds remain vowel features, and they are almost independent of speakers.

C. Vocal Tract Parameter

We use the Maeda parameters in Table I except for the seventh parameter LP. Because, when the Maeda model produces vowel sounds, the LP is steady. In the generation phase, it is possible for the Maeda parameters produced by the RNNPB to temporally fluctuate without human physical constraints. This occurs if the system does not easily associate the articulatory movements of an unexperienced 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.

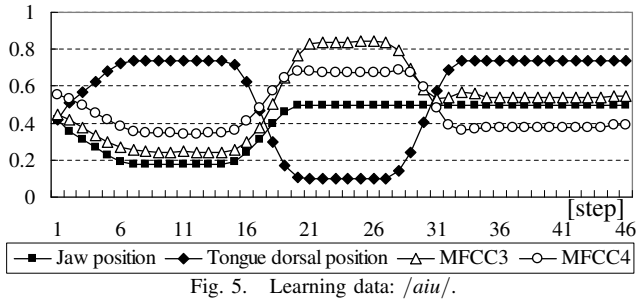

Fig. 5. Learning data: /aiu/.

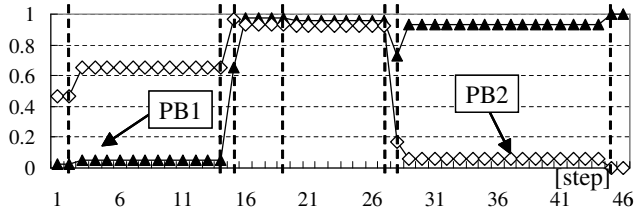

Fig. 6. The PB values of /aiu/ in the learning phase.

IV. IMITATION EXPERIMENT

We carried out a vocal imitation experiment. The organization of RNNPB is as follows: 11 input/output nodes, 40 hidden nodes, 5 context nodes, and 2 PB nodes.

In the learning phase, RNNPB learned 10 patterns of three-vowel data ($ds = 0.1$ and $N = 8$). These patterns consisted of the 5-dimensional MFCC vector and the 6-dimensional vocal tract parameters: /aiu/, /aoe/, /iue/, /iao/, /ueo/, /uia/, /eoa/, /eui/, /oai/, and /oeu/ (1350-ms and 30-ms/step), produced by the Maeda model. Learning iteration was 200,000 for each learning data, and in one iteration, the order of learning 10 patterns was alphabetical order.

In the recognition phase, we input the MFCCs of the two-vowel, three-vowel and four-vowel sounds, which are produced by one person, into the organized RNNPB, and recorded the PB values and the boundary steps for each sound. Table II lists input sound data in the recognition phase. The two-vowel data were 900-ms, the three-vowel data were 1350-ms, and the four-vowel data were 2000-ms. We set $N = 4$ in recognizing two-vowel data, and $N = 8$ in recognizing three-vowel and four-vowel data.

In the generation phase, we used the PB values and the boundary steps to reproduce each of the recorded sounds.

Figure 5 shows 4 sequences (JP and TDP of Maeda model, and the third and fourth MFCC) of the learning data /aiu/. Figure 6 shows the PB sequence for the learning data /aiu/ obtained by the organized RNNPB. The vertical dotted line represents the boundary step s_i segmented by RNNPB in the learning phase. The boundary steps, dividing the input sequence /aiu/ into flat and transition segments, in Fig. 6 were $s_1 = 2$, $s_2 = 14$, $s_3 = 15$, $s_4 = 19$, $s_5 = 27$, $s_6 = 28$ and $s_7 = 45$. We confirmed that as the size of N increases, the boundary steps become more stable in the learning phase. Similar results were also acquired for the other input data.

Figure 7 shows the PB space of the organized RNNPB. In Fig. 7, the PB values represent the phonemes of a set of

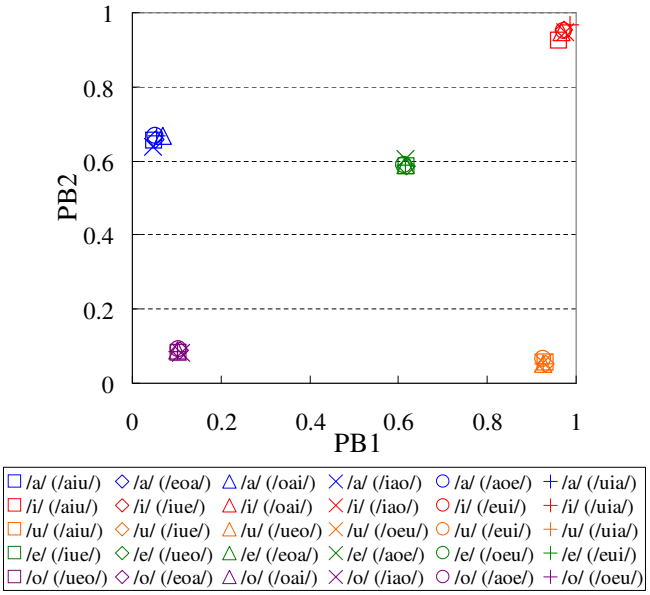

Fig. 7. The PB space in the learning phase.

three-vowel data aligned according to the length of the three longest sections of a learning sequence. The PB values for the same vowel, including the learning data, were mapped with sufficient dispersion.

Figure 8 shows the analysis of PB space. Table III shows the first and second formant (F1, F2) of vowels produced by the Maeda model. This analysis was conducted as follows:

1) The PB space was divided into 10 x 10 lattices.
2) For each lattices, each sequence of Maeda parameters was obtained through *closed loop calculation*.
3) Using the Maeda parameter sequences, 300-ms speech sounds were produced ($N = 1$).
4) The F1 and F2 averages of second half of each produced sound were calculated.
5) The square error of F1 and F2 averages from those of Table III were calculated for each vowel.
6) The vowel corresponding to the minimum square error was set at each lattice point.

In Fig. 8, each color expresses the vowel: blue is $/a/$, red is $/i/$, yellow is $/u/$, green is $/e/$, and purple is $/o/$. The clear color part is small error, and the unclear part is big error. When Fig. 7 was compared with Fig. 8, each of the PB values representing constant vowels in Fig. 7 is the clearest point in Fig. 8, and each vowel has nonlinear distribution for the F1 and F2 formants. Especially, the vowel $/a/$ is widely distributed in the PB space.

Figure 9 shows the transition of the PB values for the input data $/aeo/$ and $/eoa/$ in the recognition phase. In Fig. 9, the PB values of section S_1 for $/aeo/$ separated from those of the sections $S_{6,7}$ for $/eoa/$. These sections were longer than the other sections in the sequences. When comparing Fig. 8, we confirmed that the category of the phoneme $/a/$ in Fig. 8 corresponded to the transitions of the PB values in Fig. 9.

In the generation phase, most of the imitation sounds were similar to the original input sounds. It is confirmed that the PB values of each vowel obtained in the recognition phase correspond with those in the learning phase. Figure 10 shows

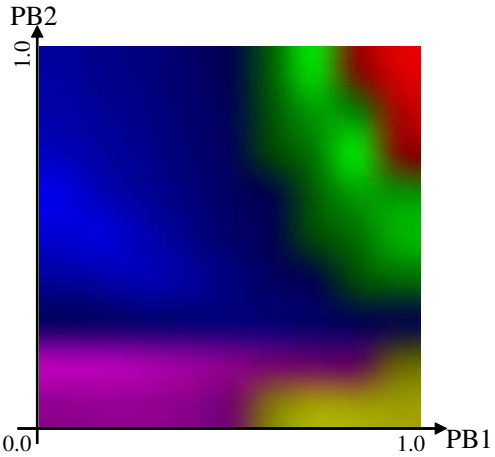

Fig. 8. Analysis of PB space.

TABLE III
THE F1 AND F2 AVERAGES OF THE MAEDA MODEL .

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

the F1-F2 mapping for each vowel of two-vowel imitated sounds. In Fig. 10, the formants of imitated sounds except for $/o/$ correspond with those of human speech sound (the map of "vowel triangle" shown in [16]).

It is confirmed that our model can imitate vocal sound involving arbitrary numbers of vowels using the vowel space in the RNNPB. The space is acquired by "babbling" of the vocal tract model with only a few sets of vowel sounds.

V. DISCUSSION

A. Vowel acquisition

Our system could encode the same vowels in acoustic signals as the near PB values in the PB space. We confirmed that there were several predictable sections of input sounds in the recognition phase, and that the sections corresponded each vowel in the self-organized PB space. In this sense, each vowel category is defined independently from the other vowels. However, in Fig. 8, it is confirmed that each vowel category is widely distributed. In Fig. 9, the transitions of PB values pass through different points in the same vowel categories. This means that the PB values representing the same vowel are changed by the adjacent vowels in a given vowel sequences. It is assumed that this represents coarticulation designed in general speech recognition systems. In this sense, each vowel is determined context dependently on the other vowels.

Tani et al. showed that the internal symbolic process was embedded in the dynamical attractor in a mobile robot system [17]. In his experiment, the robot acquired attractors representing the observed objects as activities in RNN nodes. These attractors were also represented by complex clusters, and the positions of active points were fluctuated by the context, i.e. trajectory of mobile robot. This bilateral characteristic, that is context dependency or independency, is one of the interesting and essential properties in dynamical systems representation.

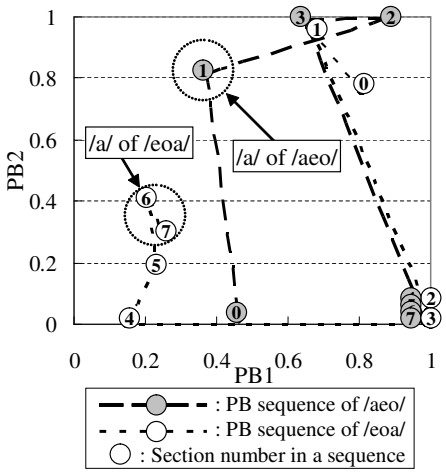

Fig. 9. PB sequences for input data /*aeo*/ and /*eoa*/ in the PB space.

B. Vowel imitation

Our system could accurately reproduce, to an extent, most of the heard sounds that were experienced or unexperienced. In the experiment, information of F0 was eliminated from input sound parameters. Due to this elimination, our system could imitate many vowel patterns of heard sounds that were experienced or unexperienced. In Fig. 9, our system could manipulate the PB values as vowels, and robustly recognize the context of sound.

The reason to fail imitate vowel /*o*/ is presumably that human infants have difficulties producing vowel /*o*/. Actually, there are large overlaps between vowel /*o*/ and the others distribution in F1-F2 space for two japanese infants [18]. This suggests that our model reflects the process of vowel acquisition.

VI. CONCLUSIONS AND FUTURE WORKS

We developed a vocal imitation system applying the segmenting method based on predictability by RNNPB. Through the experiment, the segmenting method enables our system to self-organize vowel space as the PB space without information of the number and kinds of vowels for input acoustic signals. Furthermore, imitating heard sounds, our system can manipulate the PB values as vowels in the organized PB space. For example, learning only 10 pattern of three-vowel data enables our system to imitate two-vowel and four-vowel sounds in spite of unexperienced vowel sequences. In the analysis of the organized PB space, it is confirmed that each vowel has widely distribution in the PB space and that the distribution expresses the context of speech sounds.

Our future work includes to imitate speech sounds through simulating mother and child interaction. The random babbling and consolidation learning should be introduced into our model as the exploring phase of corresponding between generated acoustic signal and articulatory movements.

VII. ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (S), and Creative Scientific Research.

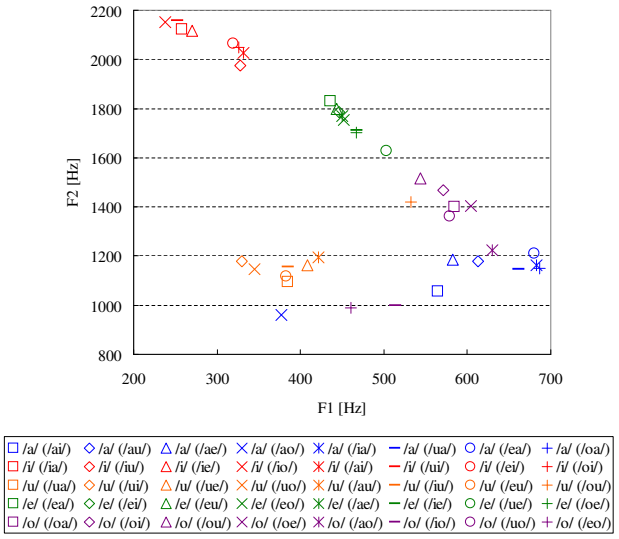

Fig. 10. The F1-F2 space in the recognition phase.

REFERENCES

[1] B. de Boer, "Self-organization in vowel systems," *J. Phonetics*, vol. 28, no. 4, pp. 441–465, 2000.
[2] P. Y. Oudeyer, "The self-organization of speech sounds," *J. Theoretical Biology*, vol. 233, no. 3, pp. 435–449, 2005.
[3] K. Miura and et al., "Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction," in *ICDL2006*, 2006.
[4] A. M. Liberman and et al., "A motor theory of speech perception," in *Proc. Speech Communication Seminar, Paper-D3*, Stockholm, 1962.
[5] L. Fadiga and et al., "Speech listening specifically modulates the excitability of tongue muscles: a tms study," *Euro. J. Cognitive Neuroscience*, vol. 15, pp. 399–402, 2002.
[6] G. Hickok, B. Buchsbaum, C. Humphries, and T. Muftuler, "Auditory-motor interaction revealed by fmri," *Area Spt. J. Cogn. Neurosci.*, vol. 15, no. 5, pp. 673–682, 2003.
[7] 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.
[8] 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.
[9] 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.
[10] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *IEEE/RSJ IROS-2007*, 2007.
[11] 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.
[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: Biological Sciences*, vol. 363, no. 1493, pp. 979–1000, 2008.
[13] N. Kitawaki and et al., "Optimum coding of transmission parameters in parcor speech analysis synthesis system," *Trans. IEICE Japan*, vol. J61-A, no. 2, pp. 119–126, 1978.
[14] M. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," in *Annu. Conf. Cog. Sci. Soc.*, 1986, pp. 513–546.
[15] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation*. Cambridge, Mass.: MIT Press, 1986.
[16] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall, 1978.
[17] J. Tani, "Model-based learning for mobile robot navigation from the dynamical systems perspective," *IEEE Trans. on SMC Part B: Cybernetics*, vol. 26, no. 3, pp. 421–436, 1996.
[18] 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.